EESTI PANK
WORKING PAPER SERIES

ARE SURVEY DATA UNDERESTIMATING THE INEQUALITY OF WEALTH?

JAANIKA MERIKÜLL, TAIRI RÕÕM

5
2019

The Working Paper is available on the Eesti Pank web site
DOI: 10.23656/25045520/052019/0167
ISBN 978-9949-606-60-3 (pdf)
Eesti Pank. Working Paper Series, ISSN 2504-5520; 5/2019 (pdf)

Are survey data underestimating the inequality of wealth?

Jaanika Meriküll and Tairi Rõõm*

Abstract

This paper uses administrative data from registers and survey data from interviews to analyse unit and item non-response in a wealth survey. It draws on the Estonian Household Finance and Consumption Survey dataset, where the survey data on income and wealth are complemented by information on the same variables from administrative sources for all the people sampled. The results show that the non-response contributes to the underestimation of wealth inequality in the survey data, as the Gini coefficient is underestimated by 6 percentage points and also the top wealth shares are substantially underestimated. The downward bias is originating from item non-response and not from unit non-response. Imputation can address the problems caused by item non-response across most of the net wealth distribution, but does not eliminate the downward bias at the top of the wealth distribution.

JEL Codes: D31 (Personal Income, Wealth, and Their Distributions); E21 (Consumption, Saving, Wealth)

Keywords: wealth distribution, unit non-response, item non-response, participation bias, wealth survey, income survey, Household Finance and Consumption Survey, Estonia

The views expressed are those of the authors and do not necessarily represent the official views of Eesti Pank or the Eurosystem.

* Authors' emails: jaanika.merikyll@eestipank.ee, tairi.room@eestipank.ee.

The authors would like to thank the participants at the Bank of Estonia research seminar, at the HFCN meeting in Krakow and at the Joint Statistical Meetings conference for their insightful comments.

Non-technical summary

This paper studies the patterns of unit and item non-response in a wealth survey and the correlations of these two types of non-response with the characteristics of the individuals surveyed. We analyse how unit and item non-response, which may potentially be selective, contribute to the estimated distribution of net wealth, and whether they create biases in such measures of inequality as the Gini coefficient, percentile ratios and top income shares.

For the analysis we use wealth data from two sources, the survey and administrative records, for the same sample of individuals. We employ the Estonian Household Finance and Consumption Survey (HFCS) dataset, which is linked to data from registers. The unique feature of these data is that the administrative records cover all net wealth components, i.e. real and financial assets and liabilities. The data on income and wealth used in this study are taken from administrative records and the information about response behaviour is obtained from the survey.

This paper has three main aims. First, we study the patterns of unit non-response. We use the administrative data on all the individuals selected for the gross sample to study the differences between participating and non-participating people. This lets us assess whether the unit non-response in the survey is random or correlated with the characteristics of the individuals sampled, and whether the distributions of wealth in the gross and net samples differ. We disentangle the response rate into two components, the contact rate and the cooperation rate, and study their patterns separately.

Second, we perform a similar analysis for item non-response. We do this using both the raw (or unedited) survey data and the administrative data. The usual approach to remedying potential biases caused by item non-response is to impute the missing data. Imputation is based on the assumption that the "missing at random" condition is satisfied, i.e. that the non-response pattern can be wholly explained by other observables collected by the survey. There are no earlier studies that test whether the "missing at random" assumption holds or how item non-response contributes to the survey bias in measures of inequality. Our analysis addresses these questions.

Third, we study the interdependence between unit and item non-response by testing whether the probability of participating is correlated with the probability of leaving survey questions unanswered. Earlier research has shown that there is a positive correlation between unit and item non-response at the individual level, since respondents who are less likely to participate are also less likely to provide answers during the interview (Yan and Curtin (2010)). This positive correlation at the individual level implies that item and unit non-response are negatively related at the aggregate level, i.e. when unit non-response increases, then more cooperative individuals are selected to participate in the survey and item non-response declines. We evaluate whether the positive correlation at the individual level exists using the Estonian Household Finance and Consumption Survey data and assess the correlation at the aggregate level by employing a cross-section of the HFCS datasets for the euro area countries.

To analyse the patterns and potential biases arising from unit and item non-response, we employ models where the dependent variable is an indicator of response. This means that we evaluate the probability of participation in the survey, or unit response, and the probability of answering the survey questions, or item response.

We employ logit regressions to study the patterns of unit response. The estimated results show that unit response is higher for contact people who are older, live in non-urban areas, have higher incomes, and have not defaulted on loan payments. Interviewer fixed effects are also relevant for unit response, and the fit of the model improves when they are added as control variables. The addition of the interviewer fixed effects has a stronger impact on the contact rate than on the cooperation rate, which suggests that the training and experience of the interviewers has less impact on response conversion than it does on success in contact. The estimated wealth inequality indicators for the cross sample are not significantly different from those that are based on the sample of respondents, which indicates that unit non-response does not bias the estimates of inequality based on the survey.

Like in the analysis of unit response, we use logit regressions to assess how item response is related with individual or household characteristics. Item response is evaluated for four variables: income, net wealth, real and financial assets, and liabilities. Income is studied at the individual level, while net wealth and its components are analysed at the household level. The estimated results imply that the response pattern is quite divergent across these variables. Item response for income is positively related with age and income. People from the capital region and in towns are more likely to answer questions about income. Item response for net wealth and its components is significantly related with fewer variables, as age, income and region do not influence it. However, what matters for the estimation of inequality is how item response for wealth-related variables correlates with wealth itself. The estimations show that it is negatively related with the level of net wealth, i.e. richer people are less likely to provide information on the values of various assets such as real estate, ownership of businesses, stocks and bonds, etc.

The negative relationship between item response and wealth causes a downward bias in the survey-based estimates of net wealth. The estimated level of net wealth is downward biased across most of the net wealth distribution and the top tail of net wealth is missing because of selective item response. We carry out an imputation of the data missing in the survey to see whether this can remedy the downward bias. Comparison of the survey statistics and inequality measures estimated on the basis of imputed data and the actual data implies that imputation can to a large extent alleviate the downward bias in the level of net wealth. It provides a good fit relative to the actual data, especially for the lower half of the distribution. However, since the top tail of the wealth distribution is missing in the survey, imputation is off support in this region and cannot reproduce the missing observations with the highest values. Therefore the estimated inequality indicators that are based on the imputed data are significantly lower than those based on the actual data. Imputation can correct the downward bias in the level of net wealth throughout most of the lower parts of its distribution, but it does not eliminate the downward bias in estimated wealth inequality.

Earlier studies have shown that survey data typically miss the top tail of the wealth distribution and therefore underestimate wealth inequality (Johansson and Klevmarken (2007), Vermeulen (2016), and Vermeulen (2018)). Our study confirms these findings and going beyond the earlier research, we are able to show what type of response causes the underestimation. Our findings indicate that unit response is not selective, i.e. it is not significantly correlated with wealth and is not a source of bias in inequality estimates, but item response is negatively related with wealth and leads to wealth inequality being underestimated in the survey.

We study the interdependence between unit and item response at the individual level. For this purpose, we use the baseline unit response model and calculate the predicted probability of participation in the survey for each reference person. Thereafter we estimate the Spearman rank correlations between the predicted probabilities of unit response and item response. Unlike the earlier research, we find that the two types of response are negatively correlated for net wealth and its components. Although the strength of the negative correlation is low, it suggests that households with a higher probability of participation have a lower probability of item response. We use the cross-sectional data of the HFCS for the euro area countries from 2013 or 2014 to evaluate the correlation between unit and item response at the aggregate level. Our estimations confirm the findings of the earlier literature and show that they are negatively related.

Contents

1. Introduction

Economic analysis of inequality is focusing increasingly on wealth (see e.g. Piketty (2014)). Both wealth and income are important determinants of individual well-being, but measuring wealth and its inequality is more challenging than measuring income and income inequality. The main reason for this is that the distribution of wealth is much more concentrated than the distribution of income (see further discussion by Cowell and Van Kerm (2015)). Just a few high-value observations can affect estimates of wealth inequality substantially, so the quality of the data has a decisive importance in studies of wealth.

Different data sources for wealth provide different insights. Information about the wealth of households can come from estate tax returns, wealth tax registers, capitalisation of incomes or wealth surveys (see e.g. Roine and Waldenström (2015); Cowell and Van Kerm (2015)). All of these sources have some shortcomings. The wealth of the deceased is not representative of that of the living; wealth tax registers are subject to errors caused by tax evasion (Roine and Waldenström (2009)); estimates based on the capitalisation of incomes overestimate inequality (Bricker et al. (2016), Lundberg and Waldenström (2018)); and survey data often miss the top tail of the wealth distribution and underestimate inequality (Vermeulen (2016) and Vermeulen (2018)). It is becoming more common to collect wealth data using surveys[1]. However, beyond the missing top tail, little is known about possible biases in wealth surveys.

The aim of this paper is to study the patterns of unit and item non-response in a wealth survey and the correlations of these two types of non-response with the characteristics of the individuals surveyed. In addition, we analyse how potentially selective unit and item non-response contribute to the estimated distribution of net wealth and whether they create biases in measures of inequality, such as the Gini coefficient, percentile ratios and top income shares.

For the analysis we use wealth data from two sources, the survey and administrative records, for the same sample of individuals. We employ the Estonian Household Finance and Consumption survey dataset, which is linked to data from registers. The unique feature of these data is that the administrative records cover all net wealth components, i.e. real and financial assets and liabilities. Having data from two sources means we can analyse the response behaviour in the survey. Although there are earlier estimates of wealth inequality that compare survey data and administrative data (e.g. Bricker et al. (2016)), there are only a few studies that compare the estimates from the two sources using data for the same set of individuals (e.g. Johansson and Klevmarken (2007)).

The current paper contributes towards filling this research gap. The contribution of the paper is threefold. First, the administrative data on all the households selected for the gross sample have been used to study the differences between participating and non-participating households. This lets us assess whether the unit non-response in the survey is random or correlated with the characteristics of the individuals sampled, and whether the distributions of wealth in the gross and net samples differ. There are some studies that use auxiliary data to analyse unit non-response, such as D´Alessio and Faiella (2002) and Pérez-Duarte et al. (2010), but none of those studies cover net wealth as completely as we do in this study. In addition to what has been done in earlier studies, we disentangle the response rate into two

[1] Examples of wealth surveys are the Bank of Italy's SHIW survey, the Federal Reserve Board's SCF, the Bank of Spain's EFF, and the ECB's HFCS.

components, the contact rate and the cooperation rate, and study their patterns separately. Our findings imply that unit non-response does not systematically bias the survey-based estimates of wealth inequality.

Second, we perform a similar analysis for item non-response. We do this using both the raw (or unedited) survey data and the administrative data. The usual approach to remedying potential biases caused by item non-response is to impute the missing data. Imputation is based on assumption that the "missing at random" condition is satisfied, i.e. that the non-response pattern can be wholly explained by other observables collected by the survey (Gelman and Hill (2006)).There are no earlier studies that test whether the "missing at random" assumption holds or how item non-response contributes to the survey bias in measures of inequality. Our analysis addresses these questions. Studies by Johansson and Klevmarken (2007), Vermeulen (2016), Vermeulen (2018) and Brzezinzki et al. (2019) show that the high-end tail of the wealth distribution is missing in wealth surveys, but they do not identify whether it is caused by unit or item non-response. We show that the missing top tail is the result of selective item non-response, whereby richer households are less likely to provide answers to questions about wealth.

Third, the interdependence between unit and item non-response is studied by testing whether the probability to participate is correlated with the probability of leaving survey questions unanswered. We test whether the response continuum model suggested by Yan and Curtin (2010) holds in our data. In this model, there is a positive correlation between unit and item non-response at the individual level, since respondents who are less likely to participate are more likely to leave questions unanswered during the interview. This positive correlation at the individual level implies that item and unit non-response are negatively related at the aggregate level, i.e. when unit non-response increases, then more cooperative individuals are selected to participate in the survey and item non-response declines. Unlike Yan and Curtin (2010), we find a negative correlation between item non-response and predicted unit non-response at the individual level.

While none of the data sources are perfect, the best source for analysing wealth inequality is presumably administrative data. These data are usually available for countries that have wealth taxes and collect the data for taxation purposes (see e.g. Lundberg and Waldenström (2018) for Sweden, Alvaredo and Saez (2010) for Spain, and Dell et al. (2007) for Switzerland). The advantage of our study is that we use wealth data from registers that are not created for taxation purposes, but for defining the legal ownership of assets or for statistical purposes. Wealth is not heavily taxed in Estonia, as there are only relatively low taxes on land while all other wealth items are untaxed. This means there are no strong tax-related incentives to misreport wealth items or to remove assets abroad. This lets us use the wealth data from administrative sources as a good measure of true value and to use these data for assessing the response behaviour in the survey.

Our findings on the pattern of unit non-response are partially in line with those of related research, but there are also some differences. Earlier papers on this topic include Bover (2011), Neri and Ranalli (2012), D'Alessio and Faiella (2002), Perez-Duarte et al. (2010) and Osier (2016). Their evidence shows that people in rural areas are more likely to participate in a survey, and our findings confirm this result. There is mixed evidence from earlier papers on such characteristics as age, education or paradata on dwelling characteristics. Our results imply that age, income and having no earlier defaults on loan payments are positively related

with the probability of participating in a survey, while dwelling appearance is negatively related.

Whether and how unit non-response affects survey-based measures of wealth inequality is dependent on its relationship with wealth. Earlier research has provided evidence either that wealth is negatively related with the probability of participation in a survey (Kennickell and Woodburn (1997), D'Alessio and Faiella (2002)), or that there is no significant relationship between wealth and unit non-response (Perez-Duarte et al. (2010)). Our analysis is in line with the results of the latter study, as we do not find that unit non-response is correlated with net wealth, nor that it contributes to the underestimation of wealth inequality in the survey.

We study item non-response for five variables: income, net wealth, real assets, financial assets, and liabilities. Age and income are positively related with whether respondents answer questions about income. Payment defaults matter for financial assets, as people who have defaulted on loan payments are less likely to reply to questions about financial assets. Relevantly for survey-based wealth inequality estimates, wealth is positively related with item non-response. We find that people with higher wealth are less likely to answers questions about real and financial assets. This selective item non-response leads to the underestimation of wealth inequality in the survey, especially for the estimated top wealth shares held by the richest 5% and 1% of people.

Our assessments imply that item non-response causes downward bias in the estimated level of net wealth throughout the wealth distribution. In addition, the top tail of the distribution is missing because of item non-response. Imputation can remedy the downward biases across most of the wealth distribution, but it cannot account for the missing top tail. While the median level of net wealth from the imputed data is close to its estimated value based on the actual data, the mean of the imputed data is lower and the inequality estimates from the imputed data are downward biased.

The paper is organised as follows: the second section presents the related literature on unit and item non-response, the third section discusses the data used, the fourth section presents the methodology, and the fifth section gives the results. The last section provides the conclusions.

2. Related literature

The existing literature on unit and item non-response is limited, mainly because of a lack of suitable data. We are not aware of any earlier papers on item non-response with a similar focus to ours. Analysis of the patterns of item non-response has to be based on the survey data with partially missing values together with data containing the actual values of the missing answers from some auxiliary source, such as administrative files. These two parallel sets of data are both available together only in very rare circumstances. The literature on unit non-response is scarce for similar reasons. Additional data besides survey responses are needed to analyse how the characteristics of survey participants differ from those of non-participants.

The main methods for unit non-response analysis are the panel attrition approach and the use of auxiliary information about participating and non-participating households. The panel attrition approach means that the characteristics of the panel households from the previous wave are used to analyse their participation in the next wave (see e.g. Neri and Ranalli (2012)). Osier (2016) describes various sources for auxiliary information on respondents and

non-respondents: ratings of dwellings collected by interviewers, the sampling frame, interviewer characteristics, administrative data, and a survey of non-respondents. The rating of dwellings and the sampling frame have been used for analysis of the unit non-response in the HFCS, the wealth survey of the euro area countries (Osier (2016)), and interviewer characteristics and a survey of non-respondents have been used for the Italian SHIW (Neri and Ranalli (2012)), while administrative or other auxiliary data have been employed for the Finnish Household Wealth Survey (Pérez-Duarte et al. (2010)) and the Italian SHIW (D´Alessio and Faiella (2002)).

The results show that the geographical region of the residence matters for the response, as households from smaller municipalities (Bover (2011), Neri and Ranalli (2012), D´Alessio and Faiella (2002)) and from outside the capital region (Pérez-Duarte et al. (2010)) are more likely to participate. The demographic variables are also relevant, as young and old people and those with a non-immigrant background are more likely to participate (Pérez-Duarte et al. (2010)). The socio-economic variables show a tendency for people with more education and those who are self-employed or work as farmers to be more likely to participate, while the response is not statistically significantly related to income (Pérez-Duarte et al. (2010)). There is evidence that households with more debt and more dividend income are more likely to participate (Pérez-Duarte et al. (2010)), but there is no conclusive evidence about the relationship between unit non-response and wealth. Kennickell and Woodburn (1999) and D´Alessio and Faiella (2002) find there to be a positive relationship between unit non-response and wealth, while Pérez-Duarte et al. (2010) find this relationship to be statistically insignificant. The unit response pattern for degree of urbanisation is similar in many studies, while the relationship between the probability of participation and some other variables such as age, education, income and wealth seems to be survey-specific or country-specific. It has been found that responses decline with income and education and that they are concave over age in the Italian SHIW (D´Alessio and Faiella (2002)). It has also been found that interviewer characteristics such as education and experience contribute to higher unit response (Neri and Ranalli (2012)).

The team conducting the HFCS has put a lot of effort into analysing the unit non-response of this survey (Pérez-Duarte et al. (2010), Osier (2016)). One of their initiatives is to collect paradata about the appearance of the home for responding and non-responding households. There are seven questions in the survey about the type of dwelling: its rating, location and outward appearance, a comparison with the neighbourhood, a rating of the surrounding buildings, and the security measures of the dwelling. Interviewers answer these questions by observing the outward appearance of the dwelling. The results show that the unit non-response patterns are country-specific, i.e. there are countries like Germany and Belgium where people from better dwellings are more likely to participate, while there are countries like Spain, Italy, Slovakia and Estonia where people from worse dwellings are more likely to participate (Bover (2011), Osier (2016)).

The closest studies to ours are the papers by D´Alessio and Faiella (2002) and Pérez-Duarte et al. (2010), which also use auxiliary information about the gross sample for their unit non-response analysis. As in their approach, we perform the conditional analysis of the response behaviour with logit regressions. We cover a richer set of auxiliary variables than these papers did. In particular, we have more comprehensive net wealth data at our disposal, covering all the conventional items of net wealth covered in surveys, except cash at home, valuables, managed accounts and private loans which together account for about 1% of net

wealth in survey-based estimates. D´Alessio and Faiella (2002) look only at financial assets and liabilities, and Pérez-Duarte et al. (2010) do not cover some financial assets such as deposits. Unlike earlier studies, we disentangle the response rate into the parts originating from success in contact and success in cooperation, and we study their patterns separately. In addition, we investigate the implications of unit non-response for the distribution of wealth in the net sample.

Item non-response is considered to be less problematic than unit non-response in wealth surveys, since potential biases arising from item non-response can usually be amended by imputation. This is an effective tool for addressing item non-response if the "missing at random" condition is satisfied. "Missing at random" implies that observations are missing conditionally on observable characteristics, i.e. the non-response pattern can be wholly explained by other observable characteristics collected by the survey (Gelman and Hill (2006)). There is, however, little evidence for whether the items actually are missing randomly. This paper fills this gap in the literature by providing estimates about the correlation of item non-response with individual characteristics. The administrative data at our disposal allows us to evaluate, which types of people and households are not answering questions about income and various wealth items, meaning we can assess how item non-response is related to income and wealth. We also perform an imputation exercise for income and wealth components where administrative and survey data are used to impute the missing items. The aim of this is to check whether the usual imputation tools can address the potential biases that arise from item non-response, so that the inequality measures for imputed data correspond to those based on the original sample.

Unit and item non-response are interlinked processes. Yan and Curtin (2010) showed that individuals whom it was harder to convince to participate in the survey had higher item non-response rates in the US Survey of Consumers. This implies a positive correlation between unit and item non-response at the individual level. Higher item non-response is also related to a lower probability of participation in the next wave of the survey among panel households. The researchers suggest a response continuum model where respondents are ordered by their propensity to participate in the survey and to respond to questions. This model suggests that the correlation between unit and item non-response turns negative at the aggregate level, so if aggregate unit non-response is increasing, more cooperative respondents are selected from the pool of respondents and item non-response declines. They confirm the negative correlation between unit and item non-response at the aggregate level, showing that increasing unit non-response is associated with declining item non-response over a period of 20 years in the US Survey of Consumers. The same regularity can be detected in the cross-section of countries conducting the HFCS, as the countries with the lowest unit non-response rates in the HFCS, such as France and Portugal, tend to have a high level of item non-response (HFCS (2017)). The correlation coefficient between the aggregate unit non-response rate and the aggregate item non-response rate for the value of household main residence is –0.38 among the countries that conducted the second wave of the HFCS in 2013–2014 (our own calculations from the HFCS 2nd wave database).

There is no earlier study that estimates the role of unit and item non-response on inequality of wealth. There is evidence that the wealthiest households are missing from the wealth surveys (Vermeulen (2016), Vermeulen (2018), Johansson and Klevmarken (2007)), but earlier papers do not disentangle the roles of unit and item non-response in this process. This would imply that survey-based estimates underestimate the inequality of wealth. We are able to

assess whether unit and item non-response contribute to the underestimation of wealth inequality in surveys by using the survey and administrative wealth data for households in Estonia.

3. Data

This paper combines two data sources, the Estonian Household Finance and Consumption Survey (HFCS) and information from administrative files. The Estonian HFCS is part of the ECB initiative to collect harmonised wealth data from euro area countries. The fieldwork for the Estonian HFCS was conducted in 2013. The unique feature of the Estonian HFCS is that the data for most of the income and wealth items were collected from two sources: from household interviews and from administrative registers. The aim of this double collection was to provide insights into the response biases in the survey and to analyse whether some parts of the interviews can be replaced by data from administrative sources in future waves.

It is not usually possible to get two parallel sets of values for various income and wealth components from different sources. When the estimates of household sector wealth are based on administrative data then there is no need to conduct a survey, so the surveys are usually only carried out when the data from other sources do not exist. In the Scandinavian countries, where data on wealth are available from administrative sources, wealth surveys are not conducted. In other European countries where wealth surveys are run, the surveys are typically not accompanied by administrative data. A few other countries participating in the Household Finance and Consumption Survey, such as Finland and Latvia, also collect data on some wealth and income items from the registers, but the coverage of wealth and income from administrative sources is not as complete as in the Estonian HFCS.

The sampling frame of the Estonian HFCS was the Population and Housing Census in 2011. People aged 18 or older were selected to participate in the survey and the participating individuals brought their households to the survey. The people that formed the gross sample for the survey are called the contact persons throughout the paper. The survey interviews consisted of two parts, with a household-level questionnaire and an individual-level questionnaire. Questions about the whole household were answered by a financially knowledgeable person (FKP), who might or might not be the contact person. Individual questions were answered by all adult family members.

Sampling design was one-stage stratified systematic sampling with 10 strata from five NUTS3 regions and two income groups. The survey design aimed to oversample wealthy households so that 20% of the sample was selected from the highest income decile and 80% from the rest of the population. The income groups were defined on the basis of the whole population of records from the Estonian Tax and Customs Board. We use the design weights throughout the paper to account for sample selection.

The survey mode was computer assisted personal interviews. The response rate was 64% in the Estonian HFCS. This relatively high response rate was mainly the result of a high cooperation rate, rather than a high contact rate (HFCS (2017)). This is a good result given that participation in the survey was not compulsory for households and that high-income households were oversampled. The eligibility rate was also quite high, which is usually a sign of a successful sampling frame (HFCS (2017)).

The administrative data were collected for all the contact persons in the gross sample, i.e. these individual-level data are available for both participating and non-participating households. We also know the household structure for participating households and the administrative data can be linked for the whole set of participating household members. Unfortunately, this cannot be done for the non-participating individuals as there is no administrative source that could provide information about their household structures. As the wealth items in the survey are collected at the household level, the analysis of item non-response for various wealth components is also done at the household level, while the analysis of unit non-response is done at the individual level. The register data encompass all the main components of income and wealth, and only a few components are not covered in this paper. The household-level income components, such as capital income and some forms of social transfers, are not covered, because the household structure is not known for all the contact persons. Private transfers, valuables and private loans are not covered by registers either. Appendix 1 provides an overview of all the income and wealth components covered.

The patterns of unit and item non-response are assessed using the information about response behaviour from the survey and deriving information on wealth and income components from the administrative data. We use the initial (or raw) survey datafile, which contains unedited and unimputed data, to get a complete understanding of the response behaviour in the interviews.

4. Methodology

4.1 Unit response

To analyse the patterns and potential biases arising from unit non-response, we employ a model that evaluates the probability of participation in the survey, or unit response. This analysis is conducted using the individual-level data from administrative sources on all eligible contact persons. The unit of analysis is a contact person who was sampled from the whole population of Estonian residents aged 18 or older. The data for responding and non-responding people were collected by a joint query from administrative records, so there are no differences in the availabilities of data for these two groups.

This paper conducts analysis of three indicators of unit response (see e.g. Osier (2016) and HFCS (2017) for the list of the other unit response indicators):

- Contact rate = People contacted / People eligible for the survey
- Cooperation rate = People interviewed / People contacted
- Response rate = People interviewed / People eligible for the survey

The response rate is the most common indicator used in the analysis of unit responses. However, the response rate itself is a combination of two factors, contact and cooperation, which can have different patterns over the characteristics of the population and can have different policy implications. Analysis of the contact rate shows the strengths and weaknesses of the sampling frame and the survey mode, while analysis of the cooperation rate highlights the segments of the population that deserve additional efforts in refusal conversion. All three indicators of unit response are regressed with a set of explanatory variables using logit regressions. The baseline model is defined as follows:

$$Response\ indicator_i^c = f(male_i, age_i, county_i, town_i, income_i, wealth_i, defaulted\ loan\ payment_i) \quad (1)$$

where i denotes individual, and c response indicator for contact rate, cooperation rate or response rate. The set of explanatory variables consists of gender, age groups, counties, a dummy variable for whether the person is living in a town, income deciles, wealth deciles and a dummy variable for whether the household had defaulted on loan payments. The variable for defaulted loan payments is derived from the credit register. This register is run by a private company and provides public information about the payment discipline of individual people and companies. There is information about unpaid bills and payments that are overdue for more than 45 days, and the database collects inputs from commercial banks, telecom companies, energy suppliers, and wholesale and retail companies.

A number of alternative specifications have been estimated in addition to the baseline model. First, the participation in various types of assets and loans has been used as an alternative to the regression with wealth deciles. This specification helps to identify the asset and liability types that are responsible for the response pattern over the distribution of net wealth. Second, the interviewer fixed effects have been introduced to test for the role of interviewers in response behaviour. Lastly, we also test for the role of the appearance of dwellings as collected by interviewers.

4.2 Item response

Like with the unit non-response analysis, we evaluate item non-response by modelling the probability of a given question being answered, i.e. item response. This analysis is conducted using the data on participating households for questions about wealth and the data on participating individuals for income-related questions. The response behaviour is assessed using unedited, raw survey data. The observation of a variable is taken to be missing if some of the components of the variable were not collected. If the variable was collected at the individual level, the household-level composite value is taken to be missing if some of the household members did not report the variable.

We analyse item response for five variables: income, net wealth, real assets, financial assets, and liabilities. All these items are collected in the survey so that first there is a question on participation in the item and then one on the value of the item. The item is taken as not reported in this paper if ownership of it was reported, but its value was not. The majority of income and wealth items did not have any missing observations in the participation variable.

Item response is analysed at the individual level for income and at the household level for net wealth and for the main net wealth components. The item response pattern is estimated with logit models where the dependent variable is a dummy variable for item response and the explanatory variables are gender, age, region, income, wealth, and a dummy for defaulted loan payments. The same set of explanatory variables is used as in the baseline model of unit response analysis, see equation (1). Gender and age refer to the household reference person in the wealth analysis at the household level. Income refers to the total income of the household, and payment default to a default by any household member in this analysis.

We have also run an imputation exercise to identify the role of item non-response in the underestimation of wealth inequality in surveys. As two separate datasets are used to analyse

the income and wealth data, with income at the individual level and wealth at the household level, the imputation is also done separately for the two datasets. First, all income components plus total household-level wealth are imputed simultaneously at the individual level. Wealth is used only as an additional covariate in this analysis and these wealth imputes are not used for the following analysis. Second, all wealth components plus total household-level income are imputed simultaneously to obtain plausible values for missing wealth items at the household level. Again, income acts as an additional covariate in this analysis and only the imputed values of the wealth components are used for the subsequent analysis. All five income[2] and 14 wealth components listed in Appendix 1 Table 2 are imputed and imputed total income is then created as the sum of the imputed income components, and imputed net wealth as the sum of the imputed assets minus the sum of the imputed liabilities.

The "mi impute" command in Stata is used for the imputation[3]. The multivariate imputation using chained equations is used, where the imputation method is predictive mean matching with five nearest neighbours. The term "chained equations" implies that all the imputation variables except the imputed one appear in the list of explanatory variables. A number of iterations are taken by estimating a sequence of univariate models starting from the equation of the variable with the fewest missing observations. The other covariates in imputation include the variables of the baseline model used to analyse the pattern of response, i.e. gender, age, region and defaulted loan payment, plus the variables that are perceived to affect income and wealth and are available for all the participants: immigration status, household size groups, level of education, and field of education. The design weights are also added to the list of covariates. The predictive mean matching implies that the replacement for the missing observation is chosen randomly from the set of observed data that has predicted values closest to the predicted value of the missing observation[4]. In our case the set of five closest values or five nearest neighbours is used.

The individual level variables like gender, age, immigration status and education refer to the household reference person in the wealth imputation at the household level. Five implicates are created and "mi estimate" commands are used to obtain standard errors for final estimates that take into account the uncertainty from imputation.

Lastly, the interdependence of unit and item response is analysed by calculating the Spearman rank correlations between the predicted values from the unit response model and the item response variables. The baseline unit non-response model gives the relationship between the explanatory variables and the unit non-response for contact persons. The coefficients of this model have been used to derive the predicted probability of unit non-response for reference persons, i.e. for the people who provided answers to the household-level questions in the survey. As the contact person in the unit non-response analysis is often not the same as the reference person, this approach provides better estimates of predicted probability than that based solely on contact persons.

---

[2] There are actually only four income components that need to be imputed, because one of the income components, private pensions, had no missing observations. See Appendix 1 Table 2 for the related statistics.

[3] The following description of methods uses the description of the commands "mi impute chained" and "mi impute pmm" in the Stata manual.

[4] As imputed values are drawn from the observed data, this method also preserves the distribution of the observed data. We have tested alternative imputation methods such as the linear regression, but this gives a very similar distribution of the imputed data to that from predictive mean matching.

5. Results

5.1 Unit response

This section presents the results of the unconditional and conditional analysis of unit response. Unconditional contact, cooperation and response rates are presented in Appendix 1 and Figure 1. The contact rate increases with income and wealth, while the cooperation rate has different patterns for income and wealth. The cooperation rate is concave over income with households from the middle income groups having higher cooperation rates than households from either of the tails of the gross income distribution. The cooperation rate decreases with net wealth, confirming the findings from the related literature that the wealthiest households are the least likely to agree to participate in wealth surveys. Since the patterns of the contact rates for income and wealth are similar, the differences in the response rate originate from the differences in the cooperation rate. The response rate is concave over income, while it is flat across the distribution of net wealth.

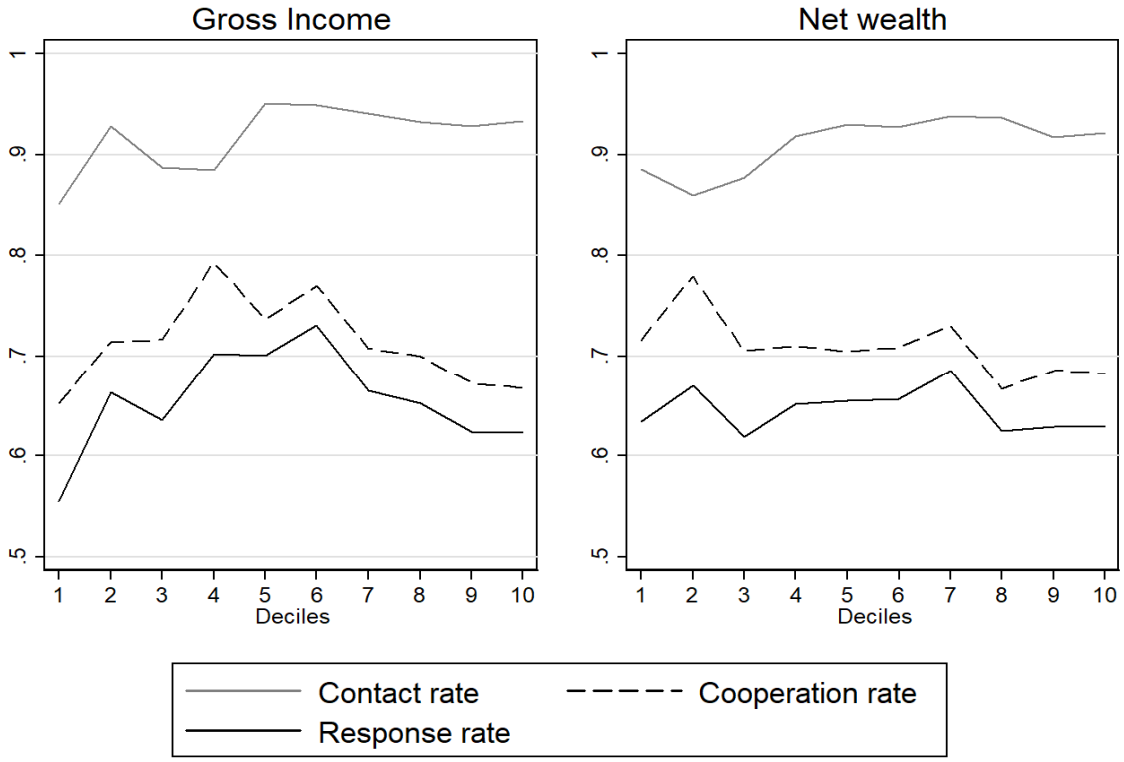

Figure 1: Unit response indicators over income and net wealth deciles
Source: Calculations of the authors from the Estonian HFCS.

Table 1 presents the results of the conditional unit response analysis. The dependent variables in the regressions are the dummy variables indicating whether it was possible to contact the given person (*contact*), whether the person contacted agreed to participate in the survey (*cooperation*), and whether that person participated in the survey (*response*). The *response* dummy equals the *contact* dummy times the *cooperation* dummy. The baseline specification

of the regression controls for gender, age groups, income deciles, wealth deciles, and a dummy for defaults on loan payments as control variables; the alternative specification uses participation in assets and liabilities as an alternative to wealth deciles. These participation variables are highly correlated with net wealth and have not been added to the same specification.

The results of the baseline estimation show that the gender of the respondent is not related to the response behaviour, but age is. Age is positively related with the probability of contact, except for the oldest age cohort (75+). The cooperation rate in contrast is higher for the 75+ cohort. In combination, these results imply that the response rate increases with age starting from the 55–64 cohort. The geographical region is strongly related to participation as the response is lower in the capital region Tallinn and in towns, and this originates from both a lower cooperation rate and a lower contact rate there.

Interestingly, the response rate is higher for higher income deciles and this is a result of the contact rate being higher. Apparently individuals with higher incomes are easier to get in contact with. In contrast to income, the relationship between net wealth and the response rate is not very strong[5]. The estimated effects tend to be negative for some net wealth deciles (the control group being the first decile), but they are only occasionally significant and so no clear pattern emerges. When the estimated effects for the response rate are negative and significant then this originates from the cooperation rate being lower.

The estimated effects for the dummy indicator for defaulting on loan payments are negative and significant for regressions on the contact rate and on the response rate. People who have unpaid bills or have missed loan payments have a lower response rate than people who have no defaulted payments, and the reason for this lower response rate is that these people are harder to contact.

The right-hand side of Table 1 presents the results for asset and debt item participation. People who have real estate assets and/or business assets have lower response rates, while people who own voluntary private pension assets and/or have mortgages have higher response rates. It is noticeable that the lower response rate among owners of real assets (real estate or business assets) is caused by their lower cooperation, meaning it is harder to convince real asset owners to participate in the survey. Similarly, the higher response rate among people who have voluntary private pensions and/or mortgages is also caused by a higher cooperation rate and not by the contact rate. The cooperation rate among people who have invested in voluntary private pensions and who have mortgages may be higher because these households are more financially literate and knowledgeable and so have lower costs from participating in a survey that demands knowledge of financial issues.

[5] This weak link between net wealth and unit response is not caused by controlling for income deciles in the regressions, as the link is still weak when the income deciles are excluded from the regression.

Table 1: Unit response indicators conditional on explanatory variables, marginal effects at averages based on logit estimation

	Baseline			**Baseline with asset/debt participation**		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
Male (base female)	–0.011	0.020	0.008	–0.010	0.036**	0.025
	(0.008)	(0.017)	(0.018)	(0.008)	(0.018)	(0.019)
Age 25–34 (base <25)	0.007	–0.078*	–0.056	0.009	–0.088**	–0.061
	(0.013)	(0.041)	(0.038)	(0.013)	(0.041)	(0.038)
Age 35–44 (base <25)	0.023*	–0.026	0.008	0.024**	–0.042	–0.003
	(0.012)	(0.039)	(0.037)	(0.012)	(0.040)	(0.038)
Age 45–54 (base <25)	0.025**	–0.010	0.022	0.027**	–0.027	0.011
	(0.012)	(0.039)	(0.038)	(0.011)	(0.040)	(0.038)
Age 55–64 (base <25)	0.038***	0.024	0.071**	0.040***	0.029	0.079**
	(0.011)	(0.037)	(0.036)	(0.011)	(0.037)	(0.035)
Age 65–74 (base <25)	0.040***	0.062	0.107***	0.043***	0.075*	0.123***
	(0.012)	(0.040)	(0.039)	(0.011)	(0.039)	(0.037)
Age >= 75 (base <25)	0.024	0.105***	0.133***	0.029**	0.111***	0.143***
	(0.015)	(0.036)	(0.038)	(0.014)	(0.036)	(0.037)
Harju county (base Tallinn)	–0.018	0.038	0.025	–0.017	0.036	0.023
	(0.016)	(0.028)	(0.030)	(0.015)	(0.028)	(0.030)
Hiiu county (base Tallinn)	0.016	0.106	0.119	0.021	0.119*	0.136*
	(0.047)	(0.071)	(0.082)	(0.042)	(0.068)	(0.078)
Ida-Viru county (base Tallinn)	0.038***	0.121***	0.154***	0.040***	0.125***	0.160***
	(0.010)	(0.023)	(0.024)	(0.009)	(0.022)	(0.024)
Jõgeva county (base Tallinn)	0.030	0.161***	0.187***	0.030	0.165***	0.192***
	(0.023)	(0.037)	(0.043)	(0.022)	(0.036)	(0.042)
Järva county (base Tallinn)	0.022	0.150***	0.168***	0.024	0.152***	0.174***
	(0.027)	(0.041)	(0.049)	(0.024)	(0.041)	(0.048)
Lääne county (base Tallinn)	0.006	0.219***	0.219***	0.003	0.217***	0.215***
	(0.026)	(0.030)	(0.040)	(0.027)	(0.030)	(0.040)
Lääne-Viru county (base Tallinn)	0.056***	0.218***	0.271***	0.055***	0.219***	0.272***
	(0.010)	(0.022)	(0.025)	(0.010)	(0.022)	(0.024)
Põlva county (base Tallinn)	–0.073*	0.040	–0.007	–0.067	0.050	0.003
	(0.044)	(0.061)	(0.065)	(0.042)	(0.061)	(0.066)
Pärnu county (base Tallinn)	–0.015	0.072**	0.062*	–0.015	0.069**	0.058*
	(0.018)	(0.031)	(0.033)	(0.018)	(0.031)	(0.033)
Rapla county (base Tallinn)	0.018	0.092*	0.109**	0.019	0.102**	0.122**
	(0.023)	(0.048)	(0.052)	(0.022)	(0.046)	(0.050)
Saare county (base Tallinn)	0.064***	0.150***	0.206***	0.063***	0.151***	0.206***
	(0.007)	(0.036)	(0.038)	(0.006)	(0.036)	(0.038)
Tartu county (base Tallinn)	–0.028*	0.063**	0.039	–0.028*	0.062**	0.039
	(0.015)	(0.025)	(0.028)	(0.015)	(0.026)	(0.028)
Valga county (base Tallinn)	0.049***	0.201***	0.246***	0.050***	0.199***	0.244***
	(0.014)	(0.030)	(0.034)	(0.013)	(0.030)	(0.034)
Viljandi county (base Tallinn)	0.043***	0.057	0.093**	0.042***	0.050	0.088**
	(0.013)	(0.040)	(0.043)	(0.013)	(0.040)	(0.042)
Võru county (base Tallinn)	0.016	0.215***	0.229***	0.014	0.216***	0.228***
	(0.025)	(0.026)	(0.035)	(0.024)	(0.025)	(0.035)
Town (base countryside)	–0.037***	–0.078***	–0.108***	–0.036***	–0.078***	–0.108***
	(0.009)	(0.022)	(0.022)	(0.009)	(0.022)	(0.022)
2nd income decile (base 1st)	0.039	0.056	0.101	0.042*	0.053	0.104
	(0.026)	(0.104)	(0.105)	(0.023)	(0.104)	(0.104)
3rd income decile (base 1st)	0.016	0.060**	0.078***	0.017	0.058*	0.076**
	(0.011)	(0.030)	(0.030)	(0.010)	(0.030)	(0.031)

	Baseline (cont.)			Baseline with asset/debt participation (cont.)		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
4th income decile (base 1st)	0.000	0.104***	0.095***	0.000	0.102***	0.094***
	(0.014)	(0.028)	(0.031)	(0.014)	(0.028)	(0.031)
5th income decile (base 1st)	0.032**	0.022	0.055	0.031**	0.017	0.049
	(0.013)	(0.037)	(0.037)	(0.012)	(0.037)	(0.038)
6th income decile (base 1st)	0.030**	0.052	0.083**	0.030**	0.048	0.079**
	(0.013)	(0.036)	(0.037)	(0.013)	(0.036)	(0.037)
7th income decile (base 1st)	0.036***	0.047	0.084***	0.035***	0.043	0.079**
	(0.010)	(0.032)	(0.032)	(0.010)	(0.032)	(0.032)
8th income decile (base 1st)	0.029***	0.040	0.074**	0.027***	0.033	0.066**
	(0.011)	(0.031)	(0.031)	(0.011)	(0.031)	(0.031)
9th income decile (base 1st)	0.027***	0.042	0.068**	0.024**	0.022	0.045
	(0.010)	(0.029)	(0.030)	(0.011)	(0.031)	(0.032)
10th income decile (base 1st)	0.034***	0.060**	0.093***	0.027***	0.024	0.052*
	(0.009)	(0.026)	(0.027)	(0.010)	(0.031)	(0.031)
2nd wealth decile (base 1st)	–0.014	–0.004	–0.019			
	(0.018)	(0.042)	(0.041)			
3rd wealth decile (base 1st)	–0.013	–0.076*	–0.075*			
	(0.019)	(0.045)	(0.043)			
4th wealth decile (base 1st)	0.006	–0.049	–0.034			
	(0.016)	(0.041)	(0.040)			
5th wealth decile (base 1st)	0.005	–0.085*	–0.076*			
	(0.017)	(0.044)	(0.043)			
6th wealth decile (base 1st)	–0.012	–0.106**	–0.104**			
	(0.020)	(0.046)	(0.044)			
7th wealth decile (base 1st)	0.002	–0.056	–0.050			
	(0.018)	(0.042)	(0.043)			
8th wealth decile (base 1st)	0.009	–0.079*	–0.062			
	(0.016)	(0.043)	(0.042)			
9th wealth decile (base 1st)	–0.011	–0.067	–0.065			
	(0.019)	(0.042)	(0.041)			
10th wealth decile (base 1st)	–0.007	–0.073*	–0.066			
	(0.019)	(0.042)	(0.041)			
Payment default (base no)	–0.120***	0.048	–0.073*	–0.116***	0.052	–0.068
	(0.029)	(0.040)	(0.041)	(0.029)	(0.040)	(0.042)
Has real estate (base no)				–0.007	–0.049**	–0.053**
				(0.009)	(0.020)	(0.021)
Has vehicles (base no)				0.002	–0.034*	–0.033
				(0.009)	(0.020)	(0.021)
Has business asset (base no)				–0.001	–0.058**	–0.056*
				(0.013)	(0.029)	(0.030)
Has deposits (base no)				0.030**	0.005	0.035
				(0.015)	(0.030)	(0.030)
Has stocks (base no) [a)]				0.031*	–0.021	0.004
				(0.017)	(0.046)	(0.047)
Has voluntary private pension (base no)				0.018	0.078***	0.091***
				(0.011)	(0.023)	(0.025)
Has mortgage (base no)				0.011	0.081***	0.089***
				(0.011)	(0.022)	(0.024)
Has other loan (base no)				0.011	0.007	0.014
				(0.009)	(0.023)	(0.024)
Has credit card debt (base no) [b)]				–0.005	0.023	0.016
				(0.010)	(0.021)	(0.022)

	Baseline (cont.)			Baseline with asset/debt participation (cont.)		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
No of observations	3475	3170	3475	3475	3170	3475
Pseudo R^2	0.104	0.068	0.073	0.108	0.073	0.078

Notes: The dependent variables in the logit regressions are 1) the dummy variable that a person could be contacted; 2) the dummy variable that a person agreed to cooperate; 3) the dummy variable that a person participated in the survey.
[a)] Also covers mutual funds and bonds. [b)] Also covers also outstanding balance of credit line/overdraft. Clustered robust standard errors in parentheses, *, **, *** refer to statistical significance at the 10, 5 and 1% levels of significance.
Source: Calculations of the authors from the Estonian HFCS.

Table 2 presents the results with the interviewer fixed effects and with the paradata about dwelling appearance. As interviewers are responsible for a particular region and town, the interviewer and county/town dummies may be perfectly collinear. To avoid the problems of perfect collinearity, the county and town dummies have been omitted from the model with interviewer fixed effects. The results show that interviewer fixed effects increase the explanatory power of the model somewhat, but leave the estimated coefficients qualitatively the same. The addition of interviewer fixed effects explains more of the model variation (i.e. results in a stronger increase of pseudo R-square) for the contact rate than for the cooperation rate. This suggests that some interviewers are more skilful at getting into contact with respondents, but it is equally difficult for all interviewers to convert refusals.

Adding the paradata about the appearance of the dwelling to the model increases pseudo R^2 to a lesser extent than adding the interviewer fixed effects did, but it can still add some explanatory power to the unit response models. The dwelling appearance information is correlated with the dummy variable of residing in a town, and this becomes statistically insignificant after controlling for the dwelling type and appearance. However, the same does not apply for the county fixed effects, which remain strongly statistically significant after controlling for dwelling location and appearance. Our results confirm the findings of an earlier study that people living in lower-ranked dwellings are more likely to participate in the Estonian HFCS (Osier (2016)). Interestingly, the dwelling rating itself and its relation to the neighbouring dwellings have opposite effects; if a dwelling is rated higher than the neighbourhood, then the person living in it has a higher probability of participating. The rest of the paradata variables, such as location, need for repair or security measures, do not have any additional explanatory power for unit response.

Table 2: Unit response indicators conditional on explanatory variables, interviewer fixed effects and paradata on dwelling, marginal effects at averages based on logit estimation

	Interviewer fixed effects			**Paradata on dwelling**		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
Male (base female)	–0.009	0.015	0.005	–0.014*	0.017	0.005
	(0.007)	(0.017)	(0.018)	(0.008)	(0.017)	(0.018)
Age 25–34 (base <25)	0.011	–0.077*	–0.045	0.010	–0.073*	–0.046
	(0.011)	(0.041)	(0.039)	(0.012)	(0.040)	(0.038)
Age 35–44 (base <25)	0.022**	–0.021	0.021	0.020*	–0.026	0.009
	(0.010)	(0.040)	(0.038)	(0.011)	(0.039)	(0.037)
Age 45–54 (base <25)	0.022**	–0.016	0.019	0.022**	–0.010	0.024
	(0.010)	(0.040)	(0.039)	(0.011)	(0.039)	(0.038)
Age 55–64 (base <25)	0.035***	0.020	0.072*	0.034***	0.028	0.075**
	(0.010)	(0.038)	(0.037)	(0.010)	(0.037)	(0.036)
Age 65–74 (base <25)	0.035***	0.050	0.100**	0.035***	0.056	0.101**
	(0.010)	(0.042)	(0.040)	(0.011)	(0.041)	(0.040)
Age >= 75 (base <25)	0.024**	0.089**	0.128***	0.025*	0.103***	0.136***
	(0.012)	(0.038)	(0.038)	(0.013)	(0.036)	(0.037)
2nd income decile (base 1st)	0.022	0.027	0.079	0.039*	0.053	0.109
	(0.032)	(0.118)	(0.110)	(0.021)	(0.105)	(0.106)
3rd income decile (base 1st)	0.013	0.050	0.070**	0.016*	0.060**	0.079***
	(0.010)	(0.031)	(0.031)	(0.010)	(0.030)	(0.030)
4th income decile (base 1st)	–0.003	0.106***	0.091***	–0.002	0.097***	0.087***
	(0.013)	(0.029)	(0.032)	(0.014)	(0.029)	(0.032)
5th income decile (base 1st)	0.029***	0.028	0.060	0.030***	0.030	0.063*
	(0.011)	(0.037)	(0.037)	(0.012)	(0.036)	(0.037)
6th income decile (base 1st)	0.025**	0.056	0.082**	0.026**	0.060*	0.090**
	(0.011)	(0.036)	(0.037)	(0.012)	(0.035)	(0.036)
7th income decile (base 1st)	0.032***	0.045	0.084***	0.033***	0.046	0.084***
	(0.009)	(0.032)	(0.032)	(0.009)	(0.031)	(0.032)
8th income decile (base 1st)	0.028***	0.058*	0.089***	0.027***	0.042	0.076**
	(0.009)	(0.030)	(0.030)	(0.010)	(0.030)	(0.031)
9th income decile (base 1st)	0.026***	0.048	0.074**	0.026***	0.051*	0.077***
	(0.009)	(0.029)	(0.031)	(0.009)	(0.029)	(0.030)
10th income decile (base 1st)	0.029***	0.055**	0.089***	0.032***	0.061**	0.094***
	(0.008)	(0.027)	(0.027)	(0.008)	(0.026)	(0.027)
2nd wealth decile (base 1st)	–0.011	0.015	–0.005	–0.008	–0.015	–0.030
	(0.016)	(0.041)	(0.041)	(0.017)	(0.044)	(0.043)
3rd wealth decile (base 1st)	–0.013	–0.072	–0.077*	–0.009	–0.065	–0.073*
	(0.017)	(0.046)	(0.044)	(0.017)	(0.045)	(0.043)
4th wealth decile (base 1st)	0.005	–0.056	–0.043	0.007	–0.048	–0.040
	(0.014)	(0.042)	(0.042)	(0.015)	(0.041)	(0.041)
5th wealth decile (base 1st)	0.007	–0.073	–0.066	0.004	–0.092**	–0.089**
	(0.014)	(0.044)	(0.044)	(0.016)	(0.044)	(0.044)
6th wealth decile (base 1st)	–0.001	–0.095**	–0.092**	–0.009	–0.109**	–0.109**
	(0.016)	(0.047)	(0.046)	(0.018)	(0.046)	(0.045)
7th wealth decile (base 1st)	0.006	–0.055	–0.049	0.001	–0.056	–0.058
	(0.015)	(0.044)	(0.044)	(0.017)	(0.042)	(0.043)
8th wealth decile (base 1st)	0.005	–0.076*	–0.066	0.008	–0.076*	–0.064
	(0.015)	(0.043)	(0.042)	(0.015)	(0.043)	(0.042)
9th wealth decile (base 1st)	–0.008	–0.064	–0.065	–0.012	–0.062	–0.070*
	(0.017)	(0.043)	(0.042)	(0.018)	(0.042)	(0.042)
10th wealth decile (base 1st)	0.003	–0.064	–0.052	–0.011	–0.074*	–0.079*
	(0.015)	(0.042)	(0.041)	(0.018)	(0.042)	(0.042)

	Interviewer fixed effects (cont.)			Paradata on dwelling (cont.)		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
Payment default (base no)	–0.106***	0.045	–0.073*	–0.113***	0.039	–0.083**
	(0.028)	(0.040)	(0.041)	–0.014*	0.017	0.005
Town (base countryside)				–0.018	–0.036	–0.049
				(0.020)	(0.045)	(0.047)
Dwelling type semi-detached house (base house)				0.000	–0.016	–0.020
				(0.024)	(0.042)	(0.044)
Dwelling type flat (base house)				–0.044***	0.029	–0.014
				(0.010)	(0.027)	(0.028)
Dwelling type other (base house)				–0.312**	–0.428***	–0.460***
				(0.143)	(0.147)	(0.095)
Rating upscale (base luxury)				0.029	0.039	0.074
				(0.018)	(0.058)	(0.058)
Rating mid-range (base luxury)				0.055*	–0.008	0.046
				(0.032)	(0.064)	(0.068)
Rating modest (base luxury)				0.046***	0.094	0.145**
				(0.015)	(0.062)	(0.062)
Rating low-income (base luxury)				0.057***	0.092	0.203***
				(0.006)	(0.103)	(0.069)
Location suburbs (base downtown)				–0.004	–0.001	–0.002
				(0.013)	(0.034)	(0.034)
Location outskirts (base downtown)				0.006	–0.017	–0.011
				(0.013)	(0.035)	(0.035)
Location isolated (base downtown)				0.008	0.011	0.022
				(0.024)	(0.054)	(0.055)
Cracks in walls (base sound)				0.003	0.027	0.027
				(0.009)	(0.020)	(0.022)
Needs repair (base sound)				0.015	0.015	0.028
				(0.014)	(0.037)	(0.039)
Dilapidated (base sound)				–0.158	0.005	–0.159
				(0.129)	(0.171)	(0.145)
As good as neighbourhood (base better)				–0.015	–0.055**	–0.068**
				(0.011)	(0.027)	(0.029)
Worse than neighbourhood (base better)				–0.068*	–0.051	–0.100**
				(0.035)	(0.050)	(0.051)
No other buildings in view (base better)				–0.017	0.098	0.089
				(0.051)	(0.060)	(0.068)
Upscale neighbourhood (base luxury)				–0.012	0.018	0.007
				(0.043)	(0.079)	(0.087)
Mid-range neighbourhood (base luxury)				–0.007	0.036	0.023
				(0.037)	(0.086)	(0.091)
Low-income neighbourhood (base luxury)				–0.027	0.002	–0.025
				(0.058)	(0.091)	(0.100)

	Interviewer fixed effects (cont.)			Paradata on dwelling (cont.)		
	Contact	Coopera-tion	Response	Contact	Coopera-tion	Response
Very low-income neighbourhood (base luxury)				–0.313	–0.016	–0.181
				(0.222)	(0.164)	(0.152)
Locked lobby (base guard)				–0.010	0.061	0.051
				(0.027)	(0.060)	(0.065)
Intercom device (base guard)				–0.003	0.007	0.004
				(0.026)	(0.061)	(0.066)
Other security (base guard)				0.010	0.077	0.086
				(0.026)	(0.057)	(0.064)
Interviewer fixed effects	yes	yes	yes	no	no	no
County fixed effects	no	no	no	yes	yes	yes
No of observations	3475	3170	3475	3475	3170	3475
Pseudo R^2	0.166	0.092	0.098	0.131	0.081	0.086

Notes: The dependent variables in the logit regressions are 1) the dummy variable that a person could be contacted; 2) the dummy variable that a person agreed to cooperate; 3) the dummy variable that a person participated in the survey. Clustered robust standard errors in parentheses, *, **, *** refer to statistical significance at the 10, 5 and 1% levels of significance.

Source: Calculations of the authors from the Estonian HFCS.

Lastly, we compare how the estimated mean and median values of income and net wealth, as well as the indicators of inequality for those two variables differ for contact persons who do or do not participate. Table 3 presents the results. These estimates are provided for four different samples: (1) the gross sample of all eligible people selected for participation in the survey; (2) the subsample of people who could not be contacted; (3) the subsample of people who refused to participate; (4) the subsample of people who participated in the survey. The estimated statistics are provided separately for income and net wealth.

The most diverse group are the people with whom no contact can be made, while the people refusing to participate have the highest levels of income and wealth. The mean income of participants is 1% higher than that of the gross sample and the mean wealth of participants is 4.7% lower than that of the gross sample. However, the difference between the gross sample of all eligible people and the sample of survey participants is not statistically significant for the median, mean or inequality measures. Appendix 2 presents the same statistics by the main net wealth components, i.e. for real and financial assets and liabilities. Nor are there any statistically significant differences for the wealth components either, with only one exception – the inequality of financial assets is weakly statistically significantly different for the eligible gross sample and for the sample of respondents. However, this difference will not show up in net wealth and we cannot conclude that the survey data underestimate wealth inequality because of unit non-response.

Table 3: Measures of income and wealth inequality among participating and not participating people

	Median	Mean	Gini	p90/p50	Top 10% share	Top 5% share	Top 1% share
Annual income, EUR							
All eligible persons (n=3475)	3841	5968	0.564	3.68	0.384	0.248	0.085
SE	40	108	0.004	0.05	0.005	0.004	0.004
No contact (n=305)	2775***	4627***	0.658***	4.33	0.457***	0.293	0.107
SE	346	359	0.017	0.37	0.022	0.023	0.022
Refused (n=950)	4002	6231	0.565	3.81	0.371	0.229*	0.066**
SE	135	208	0.008	0.11	0.008	0.007	0.005
Participated (n=3475)	3879	6043	0.550	3.58	0.384	0.248	0.088
SE	47	135	0.006	0.06	0.006	0.006	0.005
Net wealth, EUR							
All eligible persons (n=3475)	10263	45437	0.821	9.43	0.647	0.514	0.310
SE	688	5113	0.015	0.52	0.029	0.040	0.054
No contact (n=305)	2879***	48289	0.899**	29.07	0.760	0.649	0.400
SE	1642	12921	0.024	15.97	0.052	0.074	0.085
Refused (n=950)	13041	49629	0.798	8.11*	0.631	0.488	0.271
SE	1511	5133	0.014	0.55	0.025	0.032	0.033
Participated (n=3475)	10096	43316	0.818	9.46	0.638	0.501	0.298
SE	859	7429	0.024	0.57	0.048	0.065	0.090

Notes: SE refers to standard errors, analytical standard errors are shown for median and mean and bootstrapped standard errors for the Gini coefficient, p90/p50 ratio and top income shares. *, **, *** refer to statistical significance from the original sample, all eligible persons, at the 10, 5 and 1% levels of significance.

Source: Calculations of the authors from the Estonian HFCS.

The pattern of unit response is related to the context of the survey, the country of residence, and that country's institutional setting. There are some patterns that seem to repeat over various studies, for example that respondents from urban areas are less likely to participate, but for some variables such as age and income the results seem to be country and context-specific. Our study confirms the findings of an earlier study by Pérez-Duarte et al. (2010) that did not find a relationship between unit response and wealth. It seems that the response pattern is different for different wealth components. Like the study of D´Alessio and Faiella (2002), we also find that the Gini coefficients of financial assets are underestimated because of unit non-response. As financial assets make up a substantially smaller part of net wealth than real assets do, this does not cause statistically significant underestimation of net wealth inequality.

5.2 Item response

This section analyses the patterns of item response. The same baseline logit specification is used as for the unit response analysis. The dependent variable is a dummy with the value 1 when the person provided an answer and 0 when he/she did not provide an answer. The set of control variables covers gender, age, county of residence, degree of urbanisation, income and wealth deciles and a dummy for defaults on loan payments. As before, the analysis focuses on net wealth and the components of net wealth, but the results for income are also presented for comparative purposes. The income data are analysed at the individual level and the wealth data at the household level, where gender and age refer to the household reference person.

The item response rates for income, net wealth and the components of net wealth are presented in Appendix 1. The item response rates are heterogeneous across variables. The response rate is relatively high for income and liabilities at more than 80%, but it is much lower for financial and real assets, where almost every other observation has missing values. As net wealth is a combination of all the wealth components, it has the lowest item response rate at 42%.

Table 4 presents the results of the logit regressions. Comparing these figures with those presented in Table 1 indicates that the patterns of unit and item response are quite different across individual characteristics. The estimated results for different dependent variables are also quite diverse, and the pattern of item response for income is different in many ways from that for net wealth and its components. The estimated regional effects only exhibit a systematic pattern for income, but not for net wealth or its components. People outside the Tallinn capital region and in towns are less likely to give answers to income-related questions. The item response rate to questions about income is positively related to age and there is a weak concave relationship with the income level.

The regression results show that item response for real and financial assets is negatively related to net wealth, i.e. richer people are less likely to answer questions about their holdings of various real and financial asset items, such as real estate, business assets, deposits or stocks. Selective non-response to questions about wealth causes a negative correlation between the item response for net wealth and the level of wealth. As we will show later, this negative relationship causes downward bias in the estimation of wealth inequality. The response for financial assets is also negatively related to income.

The estimated regression results also show that item response depends on whether a person has defaulted on loan payments. Those who have done so are less likely to answer questions about liabilities.

Table 4: Item response patterns conditional on explanatory variables, marginal effects at averages based on logit estimation

	Income	**Net wealth**	**Real assets**	**Financial assets**	**Liabilities**
Male (base female)	0.004	–0.004	0.004	–0.007	0.017
	(0.011)	(0.024)	(0.025)	(0.025)	(0.028)
Age 25–34 (base <25)	–0.085*	–0.078	–0.087	–0.103	–0.113
	(0.044)	(0.062)	(0.093)	(0.078)	(0.135)
Age 35–44 (base <25)	–0.058	–0.121**	–0.100	–0.097	–0.132
	(0.039)	(0.059)	(0.091)	(0.077)	(0.135)
Age 45–54 (base <25)	–0.061	–0.186***	–0.167*	–0.207***	–0.154
	(0.039)	(0.056)	(0.092)	(0.075)	(0.147)
Age 55–64 (base <25)	–0.022	–0.075	–0.105	–0.122	–0.133
	(0.034)	(0.063)	(0.093)	(0.078)	(0.164)
Age 65–74 (base <25)	0.071***	–0.090	–0.144	–0.072	–0.010
	(0.023)	(0.065)	(0.096)	(0.082)	(0.142)
Age >= 75 (base <25)	0.074***	–0.068	–0.142	–0.046	–0.484*
	(0.023)	(0.067)	(0.098)	(0.084)	(0.276)
Harju county (base Tallinn)	0.029	0.178***	0.093**	0.177***	0.061
	(0.020)	(0.044)	(0.041)	(0.040)	(0.040)
Hiiu county (base Tallinn)	–0.181*	–0.089	0.077	–0.277**	–0.236
	(0.099)	(0.134)	(0.127)	(0.126)	(0.256)

	Income	Net wealth	Real assets	Financial assets	Liabilities
Ida-Viru county (base Tallinn)	–0.191***	–0.166***	–0.144***	–0.259***	–0.070
	(0.031)	(0.037)	(0.046)	(0.041)	(0.067)
Jõgeva county (base Tallinn)	0.027	0.042	0.026	0.065	0.063
	(0.043)	(0.079)	(0.078)	(0.083)	(0.072)
Järva county (base Tallinn)	0.068**	–0.063	–0.056	–0.075	0.013
	(0.034)	(0.076)	(0.090)	(0.092)	(0.115)
Lääne county (base Tallinn)	–0.011	–0.169***	–0.089	–0.216***	0.011
	(0.049)	(0.065)	(0.088)	(0.077)	(0.089)
Lääne-Viru county (base Tallinn)	–0.085**	–0.124**	–0.135**	–0.120**	–0.026
	(0.039)	(0.050)	(0.062)	(0.058)	(0.075)
Põlva county (base Tallinn)	0.067*	0.099	0.127	0.023	0.040
	(0.040)	(0.090)	(0.086)	(0.091)	(0.090)
Pärnu county (base Tallinn)	–0.117***	–0.052	–0.028	–0.143**	–0.069
	(0.037)	(0.052)	(0.054)	(0.057)	(0.068)
Rapla county (base Tallinn)	–0.067	0.119	0.028	–0.055	0.111*
	(0.058)	(0.085)	(0.082)	(0.088)	(0.061)
Saare county (base Tallinn)	–0.301***	–0.234***	–0.383***	–0.296***	0.027
	(0.055)	(0.053)	(0.060)	(0.061)	(0.075)
Tartu county (base Tallinn)	–0.204***	0.026	0.027	–0.101**	0.057
	(0.035)	(0.042)	(0.043)	(0.043)	(0.041)
Valga county (base Tallinn)	0.007	0.133*	0.194***	0.040	–0.012
	(0.038)	(0.073)	(0.062)	(0.074)	(0.126)
Viljandi county (base Tallinn)	–0.281***	0.217***	0.314***	0.015	0.059
	(0.055)	(0.068)	(0.042)	(0.071)	(0.060)
Võru county (base Tallinn)	–0.144***	–0.153**	–0.144**	–0.203***	–0.125
	(0.053)	(0.060)	(0.073)	(0.068)	(0.109)
Town (base countryside)	–0.068***	–0.010	–0.008	–0.064**	0.105**
	(0.012)	(0.031)	(0.031)	(0.030)	(0.042)
2nd income decile (base 1st)		–0.081	–0.103	–0.035	0.027
		(0.051)	(0.067)	(0.061)	(0.087)
3rd income decile (base 1st)		–0.051	–0.048	0.016	–0.040
		(0.055)	(0.065)	(0.061)	(0.111)
4th income decile (base 1st)	0.006	–0.075	–0.019	–0.014	–0.211*
	(0.023)	(0.050)	(0.060)	(0.058)	(0.109)
5th income decile (base 1st)	0.055***	–0.070	–0.014	–0.051	0.019
	(0.017)	(0.052)	(0.061)	(0.060)	(0.083)
6th income decile (base 1st)	0.034	–0.104**	–0.011	–0.100*	–0.124
	(0.021)	(0.050)	(0.059)	(0.058)	(0.107)
7th income decile (base 1st)	0.020	–0.080	–0.013	–0.103*	–0.267**
	(0.023)	(0.050)	(0.059)	(0.056)	(0.109)
8th income decile (base 1st)	0.019	–0.080	0.014	–0.089	–0.089
	(0.018)	(0.049)	(0.055)	(0.056)	(0.084)
9th income decile (base 1st)	0.011	–0.063	–0.019	–0.056	–0.115
	(0.018)	(0.048)	(0.056)	(0.054)	(0.083)
10th income decile (base 1st)	0.026	–0.160***	0.031	–0.213***	–0.114
	(0.017)	(0.043)	(0.054)	(0.052)	(0.082)
2nd wealth decile (base 1st)		0.053	0.034	0.051	–0.113
		(0.055)	(0.072)	(0.057)	(0.098)
3rd wealth decile (base 1st)	0.018	–0.063	0.053	–0.020	–0.014
	(0.032)	(0.052)	(0.066)	(0.058)	(0.061)
4th wealth decile (base 1st)	–0.005	–0.024	0.004	0.077	–0.009
	(0.027)	(0.053)	(0.067)	(0.055)	(0.060)
5th wealth decile (base 1st)	–0.009	–0.160***	–0.058	–0.065	–0.080
	(0.024)	(0.044)	(0.069)	(0.057)	(0.077)

	Income	Net wealth	Real assets	Financial assets	Liabilities
6th wealth decile (base 1st)	–0.006	–0.133***	–0.109	–0.033	–0.041
	(0.022)	(0.046)	(0.069)	(0.057)	(0.076)
7th wealth decile (base 1st)	–0.008	–0.165***	–0.120*	–0.088	–0.097
	(0.022)	(0.044)	(0.069)	(0.057)	(0.074)
8th wealth decile (base 1st)	–0.020	–0.167***	–0.118*	–0.158***	0.034
	(0.022)	(0.044)	(0.069)	(0.057)	(0.055)
9th wealth decile (base 1st)	–0.014	–0.124***	–0.095	–0.094	0.015
	(0.022)	(0.048)	(0.070)	(0.058)	(0.058)
10th wealth decile (base 1st)	–0.059**	–0.236***	–0.261***	–0.137**	–0.032
	(0.027)	(0.039)	(0.065)	(0.058)	(0.064)
Payment default (base no)	0.024	–0.026	–0.147	0.112	–0.145*
	(0.023)	(0.083)	(0.099)	(0.098)	(0.076)
No of observations	3789	2198	1986	2124	880
Pseudo R^2	0.113	0.078	0.068	0.083	0.079

Notes: Dependent variable: dummy of item response, 0 – did not respond, 1 – responded. Clustered robust standard errors in parentheses, *, **, *** refer to statistical significance at the 10, 5 and 1% levels of significance.
Source: Calculations of the authors from the Estonian HFCS.

Lastly, we present evidence on whether the mean and median values for income and net wealth and the inequality indicators for these variables are different for those who report a given item and for those who do not report it (see Table 5). These statistics are given for four different samples: (1) the sample of all the people participating in the survey, with the actual values of the variables; (2) the subsample of people who did not answer the questions about income or wealth ("non-reporters"); (3) the subsample of people who answered questions about income or wealth ("reporters"); and (4) the sample of all the people participating in the survey, with imputed values for missing observations.

Table 5: Measures of income and wealth inequality among reporters and non-reporters

	Median	Mean	Gini	p90/p50	Top 10% share	Top 5% share	Top 1% share
Annual income, EUR							
All persons (n=3800)	4014	6484	0.517	3.58	0.367	0.238	0.082
SE	40	114	0.004	0.06	0.005	0.004	0.004
Non-reporters (n=599)	4493	6971	0.559***	3.62	0.375	0.239	0.064*
SE	299	316	0.010	0.18	0.010	0.010	0.006
Reporters (n=3201)	3979	6392	0.507	3.53	0.366	0.237	0.084
SE	36	121	0.005	0.06	0.005	0.005	0.004
All persons, imputed (n=3800)	4020	6472	0.513	3.58	0.363	0.234	0.082
SE	41	117	0.004	0.05	0.004	0.004	0.004
Net wealth, EUR							
All households (n=2198)	33641	76659	0.699	4.79	0.536	0.412	0.228
SE	1445	7992	0.023	0.13	0.035	0.044	0.055
Non-reporters (n=1279)	41012**	94109	0.698	4.33	0.556	0.440	0.266
SE	1941	13619	0.033	0.18	0.048	0.059	0.074
Reporters (n=919)	24230***	52610**	0.680	5.82**	0.481	0.328	0.125
SE	1970	2822	0.009	0.33	0.013	0.014	0.009
All households, imputed (n=2198)	35400	65079	0.638**	4.37*	0.447**	0.306**	0.115*
SE	1765	2494	0.006	0.12	0.007	0.007	0.004

Notes: SE refers to standard errors, analytical standard errors are shown for the median and mean and bootstrapped standard errors for the Gini coefficient, p90/p50 ratio and top income shares. *, **, *** refer to statistical significance from the original sample, all eligible persons, at the 10, 5 and 1% levels of significance.
Source: Calculations of the authors from the Estonian HFCS.

The figures presented in Table 5 indicate that item non-response does not cause significant biases in the estimated level or inequality indicators for income, but the same does not hold for net wealth. The median and mean values of net wealth are lower for the subsample of reporters and higher for the subsample of non-reporters than for the sample of all the people participating. This implies that item non-response causes downward bias in the estimated level of wealth in the survey.

To see how prevalent this bias is throughout the wealth distribution, we estimated Kernel density functions of net wealth for the sample of all those participating and for the subsample of those who answered all the questions about wealth. These estimated Kernel densities are presented in the upper graph of Figure 2. This graph shows that the estimated distribution of net wealth for the subsample of reporters is shifted to the left relative to the distribution for all the survey participants. The leftward shift can be observed for most of the net wealth distribution, indicating that item non-response causes downward bias in the estimated level of net wealth throughout the net wealth distribution. In addition, the top tail of the distribution for the full sample of all the survey participants is longer than that for the subsample of reporters. This implies that the richest households are not covered by the survey because of selective item non-response. The maximum level of net wealth for households that report this variable in the survey is 1.1 million euros, while the maximum level of net wealth for households that do not report net wealth in the survey is 13.6 million euros.

To understand whether imputation can alleviate the downward bias in the survey-based estimates of net wealth caused by selective item non-response, we evaluated the sample statistics and the inequality indicators for the sample of all the survey participants where the actual observations were replaced by imputed values for the missing answers in the survey. These estimates are given at the bottom of Table 5. The median level of net wealth for the imputed data is close to that for the actual data, so it seems that imputation can correct for downward bias at least in the lower half of the wealth distribution. However, the estimated mean level of wealth and the inequality measures that are based on imputed data are significantly lower than the estimates that are based on the actual data. These differences are especially sizeable for the estimated shares of wealth held by the top 5% and 1% of households.

To assess the impact of imputation further, we estimated Kernel density functions for the sample of all the survey participants with the actual data and for the sample of the same people with the imputed data. These density functions are shown in the lower graph of Figure 2. It can be seen that the two density functions practically coincide for the lower half of the wealth distribution and are also close to each other for most of the upper half. Evidently imputation can remedy the downward biases caused by selective item non-response for most of the distribution, and it works especially well in replicating the actual distribution for lower values of wealth up to the median. However, the top tail of the distribution based on the actual data is much longer than that based on the imputed data. The maximum value of net wealth in the imputed data is close to that in the reported data, i.e. 1.1 million euros. This implies that the wealthiest households are off support in the imputation exercise and this is why the imputation results in the underestimation of the net worth of the wealthiest and in the underestimation of wealth inequality. The main source of the missing tail is self-employment business assets, as the mean of the reported observations for this asset class is substantially lower than the mean for the observations that are not reported.

Appendix 2 provides the same analysis for real and financial assets and liabilities. It is evident that the underestimation of mean wealth and wealth inequality resulting from item

non-response originates from real and financial assets and not from liabilities. These results are in line with the findings from the logit regressions on item response presented in Table 4, which showed that wealthier households are less likely to report their real and financial assets, while there is no statistically significant relationship between wealth and item non-response for liabilities.

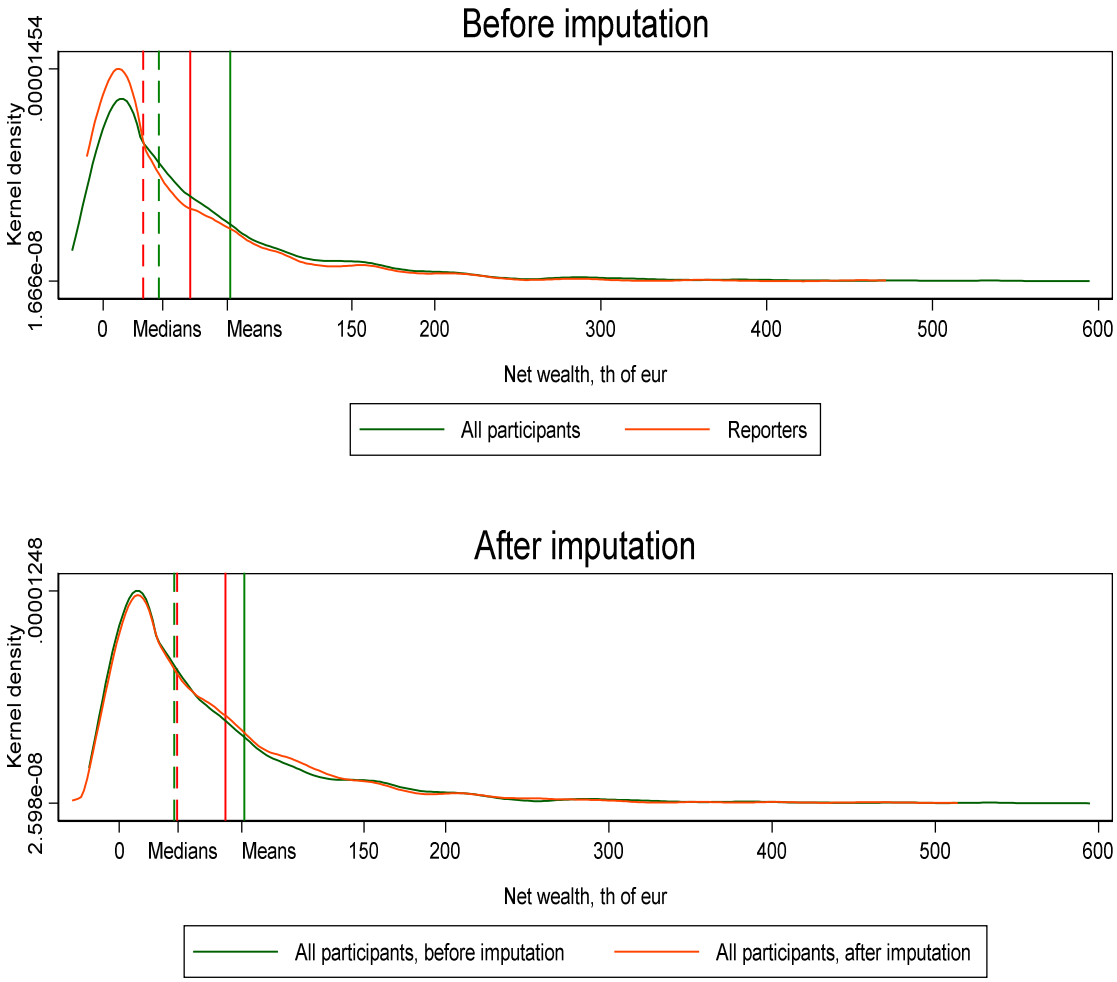

Figure 2: The distribution of net wealth before and after imputation

Notes: The bottom and top 1% of the distribution are omitted from the figure in order to achieve clarity of distributions.

Source: Calculations of the authors from the Estonian HFCS.

The study by Brzezinzki et al. (2019) also shows that wealth inequality is likely to be underestimated in the survey data of the Estonian HFCS. They impute the missing top tail of the wealth distribution using the publicly available list of the top 500 richest people. They find the underestimation to be of the same magnitude that we do. However, they do not have the whole sample of registry data at their disposal and they cannot disentangle where the underestimation is coming from, i.e. whether it is caused by unit or item non-response.

To conclude, the regression results presented in this subsection showed that wealthier people are less likely to answer questions about wealth. This negative relationship between item response and wealth has consequences for the survey-based wealth estimates, as both the estimated level of wealth and wealth inequality are downward biased because of the selective item response. Imputation can correct the bias in the level of net wealth throughout most of its distribution, but does not eliminate the downward bias in the estimated wealth inequality.

5.3 The interdependence of item and unit response

The related literature suggests that there is interdependence between unit and item response (see e.g. Yan and Curtin (2010)). Our data allow us to test directly at the individual level whether they are related. For this purpose, we use the baseline unit response model and calculate the predicted probability of participation in the survey for each reference person. The related literature suggests that people whom it is hard to convince to participate in the survey, i.e. people with low predicted probability of unit response, will leave more questions unanswered during the interview (Yan and Curtin (2010)).

Table 6 presents the Spearman rank correlations between predicted probability of unit response and item response for the income and wealth variables analysed. The findings do not confirm the prediction from the earlier literature. There is a positive but weak and statistically insignificant correlation between unit and item response for income, while the correlation between unit and item response is negative and statistically significant for net wealth and its components. Although the strength of the negative correlation is low, it suggests that households with a higher probability of participation actually have a lower probability of item response.

Table 6: Correlation between predicted unit response and item response, Spearman rank correlation

	Predicted unit response	**Item response of net wealth**	**Item response of real assets**	**Item response of financial assets**
Item response on income (n=3800)	0.024			
Item response on net wealth (n=2198)	–0.039*			
Item response on real assets (n=1986)	–0.047**	0.637***		
Item response on financial assets (n=2124)	–0.037*	0.727***	0.246***	
Item response on liabilities (n=880)	–0.113***	0.368***	0.149***	0.143***

Notes: *, **, *** refer to statistical significance at the 10, 5 and 1% levels of significance.

Source: Calculations of the authors from the Estonian HFCS

6. Conclusions

The aim of this paper is to analyse, which households participate in household wealth surveys, which give answers to the interview questions, and how this response behaviour affects the estimates of income and wealth inequality. We use the Estonian HFCS dataset, where the survey data are linked with register data for the same individuals. The data on income and wealth used in this study are taken from administrative records and the information about

response behaviour is obtained from the survey. The administrative data are available for the whole gross sample of contact persons, i.e. for those participating in the survey and for those not participating. The regression results show that unit response is higher for contact persons who are older, live in non-urban areas, have higher incomes, and have not defaulted on loan payments. Interviewer fixed effects are also relevant for unit response, and the fit of the model improves when they are added as control variables. The addition of the interviewer fixed effects has a stronger impact on the contact rate than on the cooperation rate, which suggests that the response conversion is more difficult to achieve through the training and experience of the interviewers than the success in contact is. The estimated wealth inequality indicators for the cross sample are not significantly different from those that are based on the sample of respondents, which indicates that unit non-response does not bias the estimates of inequality based on the survey.

Item response is evaluated for four variables: income, net wealth, real and financial assets, and liabilities. The estimated results imply that the response pattern is quite divergent across these variables. Item response for income is positively related with age and income. People from the capital region and in towns are more likely to answer questions about income. The item response for net wealth and its components is significantly related with fewer variables, as age, income and region do not influence it. However, what matters for the estimation of inequality is how item response for wealth-related variables correlates with wealth itself. The estimations show that it is negatively related with the level of net wealth, i.e. richer people are less likely to provide information on the values of various assets such as real estate, ownership of businesses, stocks and bonds, etc.

The negative relationship between item response and wealth causes a downward bias in the survey-based estimates of net wealth. The estimated level of net wealth is downward biased across most of the net wealth distribution and the top tail of net wealth is missing because of selective item response. We carry out an imputation of the data missing in the survey to see whether this can remedy the downward bias. Comparison of the survey statistics and inequality measures estimated on the basis of imputed data and the actual data implies that imputation can to a large extent alleviate the downward bias in the level of net wealth. It provides a good fit relative to the actual data, especially for the lower half of the distribution. However, since the top tail of the wealth distribution is missing in the survey, imputation is off support in this region and cannot reproduce the missing observations with the highest values. Therefore the estimated inequality indicators that are based on imputed data are significantly lower than those based on the actual data. Imputation can correct the downward bias in the level of net wealth throughout most of the lower parts of its distribution, but it does not eliminate the downward bias in estimated wealth inequality.

Earlier studies have shown that survey data typically miss the top tail of the wealth distribution and therefore underestimate wealth inequality (Johansson and Klevmarken (2007), Vermeulen (2016) and Vermeulen (2018)). Our study confirms these findings and going beyond the earlier research, we are able to show what type of response causes the underestimation. Our findings indicate that unit response is not selective, i.e. it is not significantly correlated with wealth and is not a source of bias in inequality estimates, but item response is negatively related with wealth and leads to wealth inequality being underestimated in the survey.

It remains to be seen to what extent the findings in this study apply for other countries as well. The result that unit non-response is not related to wealth is also confirmed by an earlier

study on Finnish data (Perez-Duarte et al. (2010)), but is not supported by the US data, where wealth and the probability of responding to the wealth survey are positively related (Kennickell and Woodburn (1999)). Further research on these topics is warranted to see whether the results presented here also apply in the wider context.

References

Alvaredo, F.; Saez, E. (2010): Income and Wealth Concentration in Spain in a Historical and Fiscal Perspective. In Atkinson, A, B., Piketty, T. (Eds.), Top incomes global perspective. Oxford University Press.

Bricker, J.; Henriques, A.; Krimmel, J.: Sabelhaus, J. (2016): Measuring Income and Wealth at the Top Using Administrative and Survey Data. *Brookings Papers on Economic Activity*, Spring 2016, pp. 261–312.

Bover, O. (2011): The Spanish survey of household finance (EFF): Description and methods of the 2008 wave. Documentos Ocasionales no 1103, Banco De España.

Brzezinzki, M.; Salach, K.; Wronski, M. (2019): Wealth inequality in Central and Eastern Europe: Evidence from joined household survey and rich lists' data. University of Warsaw Faculty of Economic Sciences Working Papers, No 9/2019.

Cowell, F., A.; Van Kerm, P. (2015): Wealth inequality: A survey. *Journal of Economic Surveys*, 29(4), pp. 671–710.

D´Alessio, G.; Faiella, I. (2002): Non-response behaviour in the Bank of Italy's Survey of Household Income and Wealth. Banca D'Italia Working Paper no 462.

Dell, F.; Piketty, T.; Saez, E. (2007): Income and wealth concentration in Switzerland over the 20th century. In Atkinson, A, B., Piketty, T. (Eds.), Top incomes over the 20th century, Oxford University Press.

HFCS (2013): The Eurosystem Household Finance and Consumption Survey: Results from the first wave. ECB Statistics Paper Series no 2.

HFCS (2017): The Household Finance and Consumption Survey: Methodological report for the second wave. ECB Statistics Paper Series no 17.

Gelman, A.; Hill, J. (2006): Data Analysis Using Regression and Multilevel/Hierarchical Models. Columbia University Press, New York.

Johansson, F.; Klevmarken, A. (2007): Comparing register and survey wealth data. Department of Economics, Uppsala University.

Kennickell, A. B.; Woodburn, R. L. (1999): Consistent Weight Design for the 1989 and 1995 SCFs, and the Distribution of Wealth. *Review of Income and Wealth*, 45(2), pp. 193–215.

Lundberg, J.; Waldenström, D. (2018): Wealth inequality in Sweden: What can we learn from capitalized income tax data? *The Review of Income and Wealth*, 64(3), pp. 517–541.

Neri, A.; Ranalli, M. G. (2012): To misreport or not to report? The measurement of household financial wealth. Banca D'Italia Working Paper no 870.

Osier, G. (2016): Unit non-response in household wealth surveys. ECB Statistics Paper Series no 15.

Pérez-Duarte, S.; Sánchez-Muñoz, C.; Törmälehto, V.-M. (2010): Re-weighting to reduce unit non-response bias in household wealth surveys: a cross-country comparative perspective

illustrated by a case study. Proceedings of the European Conference on Quality in Official Statistics.

Piketty, T. (2014): Capital in the Twenty-First Century. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Roine; J.; Waldenström, D. (2009): Wealth Concentration over the Path of Development: Sweden, 1873–2006. *The Scandinavian Journal of Economics*, 111(1), pp. 151–187.

Roine; J.; Waldenström, D. (2015): Long-Run Trends in the Distribution of Income and Wealth. In Atkinson, A., B., Bourgiognon; F. (Eds.), Handbook of Income Distribution, Vol. 2, North-Holland, Amsterdam.

Vermeulen, P. (2016): Estimating the Top Tail of the Wealth Distribution. *American Economic Review: Papers & Proceedings*, 106(5), pp. 646–650.

Vermeulen, P. (2018): How fat is the top tail of the wealth distribution? *The Review of Income and Wealth*, 64(2), pp. 357–387.

Yan, T.; Curtin, R. (2010) The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research*, 22(4), pp. 535–551.

Appendix 1. Descriptive statistics of unit and item response rates

Table A1.1. Unit response rates

	Participation	Cooperation	Response
Average rate	0.912	0.708	0.645
Number of observations	3 475	3 170	3 475

Source: Calculations of the authors from the Estonian HFCS.

Table A1.2. Item response rates of income and wealth components

	Item response rate	Number of observations
Total income	0.842	3 800
…wage income	0.802	2 639
…self-employment income	0.667	139
…public pensions	0.945	1 362
…private pensions	1.000	36
…unemployment benefits	0.789	111
Total real assets	0.602	1 986
...household main residence	0.688	1 778
...other real estate	0.613	844
...vehicles	0.877	1 157
...self-employment business	0.489	313
Total financial assets	0.571	2 124
...deposits	0.610	2 113
...mutual funds	0.702	64
...bonds	0.353	4
...shares	0.810	55
...private pensions	0.496	377
Total liabilities	0.815	880
...main residence mortgages	0.824	509
...other mortgages	0.830	99
...non-mortgage loans	0.851	316
...credit line/overdraft	0.891	64
...credit card debt	0.903	212
Net wealth	0.421	2 198

Source: Calculations of the authors from the Estonian HFCS.

Appendix 2. The measures of inequality by the components of wealth

Table A2.1. Unit non-response and the measures of inequality by the components of wealth

	Median	Mean	Gini	p90/p50	Top 10% share	Top 5% share	Top 1% share
Real assets, thousand EUR							
All eligible persons (n=3475)	11109	45839	0.775	8.80	0.614	0.485	0.292
SE	880	5118	0.018	0.49	0.032	0.042	0.055
No contact (n=305)	1875***	48017	0.840	44.60*	0.713	0.605	0.334
SE	2080	12303	0.030	19.79	0.057	0.077	0.079
Refused (n=950)	15636	50471	0.746	7.34	0.582	0.448	0.251
SE	2231	4968	0.016	0.74	0.027	0.034	0.035
Participated (n=3475)	10105	43628	0.777	9.44	0.615	0.483	0.290
SE	1032	7475	0.029	0.56	0.051	0.068	0.091
Financial assets, thousand EUR							
All eligible persons (n=3475)	378	5089	0.856	28.54	0.754	0.610	0.346
SE	36	421	0.008	1.83	0.014	0.021	0.033
No contact (n=305)	100***	6504	0.938*	72.63	0.910**	0.806	0.592
SE	44	3424	0.035	27.14	0.056	0.118	0.232
Refused (n=950)	341	5570	0.870	32.70	0.783	0.638	0.367
SE	70	762	0.011	4.24	0.020	0.031	0.044
Participated (n=3475)	466	4697	0.832*	23.15	0.712	0.556	0.282
SE	52	329	0.007	1.78	0.013	0.018	0.024
Liabilities, thousand EUR							
All eligible persons (n=3475)	0	5492	0.900	.	0.851	0.628	0.247
SE	0	277	0.003	.	0.007	0.011	0.010
No contact (n=305)	0	6233	0.904	.	0.876	0.627	0.217
SE	12	1233	0.010	.	0.023	0.044	0.046
Refused (n=950)	0	6412	0.890	.	0.826	0.596	0.211
SE	0	580	0.005	.	0.014	0.020	0.014
Participated (n=3475)	0	5010	0.902	.	0.861	0.633	0.250
SE	0	313	0.003	.	0.010	0.014	0.014

Notes: SE refers to standard errors, analytical standard errors are shown for the median and mean and bootstrapped standard errors for the Gini coefficient, p90/p50 ratio and top income shares. *, **, *** refer to statistical significance from the original sample, all eligible persons, at the 10, 5 and 1% levels of significance.

Source: Calculations of the authors from the Estonian HFCS.

Table A2.2. Item non-response and the measures of inequality by the components of wealth

	Median	Mean	Gini	p90/p50	Top 10% share	Top 5% share	Top 1% share
Real assets, EUR							
All persons (n=1986)	42586	85595	0.640	4.00	0.504	0.389	0.220
SE	1515	8981	0.027	0.11	0.038	0.046	0.056
Non-reporters (n=775)	48831	116616	0.679	4.07	0.565	0.463	0.307
SE	2756	22126	0.046	0.19	0.063	0.077	0.095
Reporters (n=1211)	37491	65068*	0.586	3.93	0.424	0.291*	0.112*
SE	1878	2754	0.009	0.15	0.011	0.012	0.007
All persons, imputed (n=1986)	45172	74174	0.568**	3.70	0.410**	0.280**	0.103**
SE	1699	2698	0.006	0.09	0.007	0.007	0.004
Financial assets, EUR							
All households (n=2124)	1656	8228	0.777	12.19	0.633	0.472	0.206
SE	141	444	0.006	0.81	0.011	0.014	0.014
Non-reporters (n=916)	3115***	10687**	0.747**	8.78**	0.608	0.449	0.183
SE	260	755	0.009	0.64	0.015	0.017	0.015
Reporters (n=1208)	931***	6381*	0.795	17.43**	0.647	0.484	0.231
SE	115	525	0.009	1.28	0.017	0.022	0.025
All households, imputed (n=2124)	1254	6611*	0.776	13.12	0.625	0.467	0.222
SE	139	429	0.007	0.60	0.012	0.015	0.016
Liabilities, EUR							
All persons (n=880)	4 055	20 823	0.726	15.04	0.517	0.339	0.109
SE	656	1 145	0.007	1.23	0.012	0.011	0.008
Non-reporters (n=151)	6 106	24 293	0.709	12.24	0.542	0.362	0.167
SE	2 503	3 207	0.017	3.49	0.026	0.027	0.032
Reporters (n=729)	3 888	20 034	0.729	15.57	0.515	0.336	0.118
SE	704	1 199	0.008	1.26	0.014	0.013	0.008
All persons, imputed (n=880)	4 277	20 400	0.719	14.33	0.500	0.327	0.107
SE	709	1 242	0.007	1.06	0.012	0.010	0.007

Notes: SE refers to standard errors, analytical standard errors are shown for the median and mean and bootstrapped standard errors for the Gini coefficient, p90/p50 ratio and top income shares. *, **, *** refer to statistical significance from the original sample, all eligible persons, at the 10, 5 and 1% levels of significance.

Source: Calculations of the authors from the Estonian HFCS.

Working Papers of Eesti Pank 2019

No 1
Jacopo Bonchi. Asset price bubbles with low interest rates: not all bubbles are alike

No 2
Thomas Y. Mathä, Stephen Millard, Tairi Rõõm, Ladislav Wintr and Robert Wyszynski. Shocks and labour cost adjustment: evidence from a survey of European firms

No 3
Juan Carlos Cuestas, Yannick Lucotte, Nicolas Reigl. The evolution and heterogeneity of credit procyclicality in Central and Eastern Europe

No 4
Jaanika Meriküll, Merike Kukk, Tairi Rõõm. What explains the gender gap in wealth? Evidence from administrative data