# Revision of the cross-sectional and longitudinal auxiliary vectors in the Swedish SILC

This report was funded by the European Union.

2021
Living conditions 2021:5

# Revision of the cross-sectional and longitudinal auxiliary vectors in the Swedish SILC

Producer
Statistics Sweden, Social statistics and analysis
SE-171 54 Solna, Sweden
+46 10-479 40 00

Enquiries
Jannis Kalpouzos
+46 01-479 46 54
jannis.kalpouzos@scb.se
Jens Malmros
+46 10-479 44 25
jens.malmros@scb.se



ISSN: 1654-1707(Online)
URN:NBN:SE:SCB-2021-LEBR2105_pdf

This publication is only available in electronic form on www.scb.se

# Preface

From 2021, the Swedish SILC and LCS will have a new design. The new IESS precision requirements, the need for further integration between the Swedish SILC and the Swedish LCS, and general improvements to the efficiency and quality of the Swedish living conditions statistics were important aspects in the development of the new design. A rotating panel survey will cover the material of both surveys, the number of panels will increase from four to six, and panel samples will have a new sampling design.

Statistics Sweden developed the new design through several project initiatives on overall design, sampling, and estimation. The present work on choosing an auxiliary vector for the new design, which efficiently corrects for non-response bias and reduces the variance of estimates, marks one of the final efforts in this work.

This report was funded by the European Union. The content of this report represents the views of the authors only and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

Statistics Sweden December 2021

Thomas Helgeson
Head of section

Marie Lidéus
Head of unit, Population and living conditions

# Table of contents

# Abstract

From 2021, the Swedish Survey of Income and Living Conditions will have a new design, in which the number of panels increases from four to six, the cross-sectional sample size increases from 11 600 to 20 000, and the stratification variable changes from age categories to NUTS2 regions. The new design also features further integration with the Swedish Living Conditions Survey, in particular by using a joint estimation procedure. These changes motivate a revision of the cross-sectional auxiliary vector, which is the purpose of the present work.

We evaluate the set of possible auxiliary variables through descriptive analysis, regression analysis, indicator analysis, and estimation of survey variables and register variables. We select auxiliary variables and auxiliary vectors sequentially throughout these analyses. The result of this work is a revised auxiliary vector for the cross-sectional estimation procedure of the Swedish SILC. We also introduce and evaluate calibration estimation for the longitudinal estimation procedure.

# 1 Introduction

From 2021, the Swedish Survey of Income and Living Conditions (SILC) will undergo a major redesign. The number of panels will increase from four to six, and the sample size will increase from 11 600 to 20 000[1]. The panel sample design will change with respect to stratification, allocation, and sampling procedure. The redesign facilitates further integration between SILC and the Swedish Living Conditions Survey (LCS). Prior to 2021, SILC and LCS were standalone surveys with only partial integration through minor sample and content overlap. The new design features further integration of content and sample for all survey variables. See (SCB, 2018) and (SCB, 2020) for further details on the new design.

Statistics Sweden introduced calibration estimation in the cross-sectional estimation procedure of the Swedish SILC in 2016.[2]. Calibration estimation adjusts for non-response bias, reduces the variance of estimates, and reproduces population distributions. We utilize register variables such as age, sex, education level, and income in the calibration procedure. These variables are the *auxiliary variables* used in the survey, and together, they form an *auxiliary vector*. Because of the new design, we need to revise the cross-sectional auxiliary vector.

We choose not to limit ourselves to merely adapting the auxiliary vector to the new design; rather, we will evaluate different categorizations of the current auxiliary variables and introduce register variables currently not included in the auxiliary vector. The previous auxiliary vectors for LCS and SILC are not identical. Because the new design features only one auxiliary vector, it should consider both SILC and LCS needs.

The present work also includes a revision of the longitudinal estimation scheme. Currently, the auxiliary information in the longitudinal estimation procedure is limited to stratification variables. We will introduce calibration estimation for longitudinal estimates. The choice of auxiliary variables will set out from the new cross-sectional auxiliary vector.

In what follows, we assume that the reader has a working knowledge of SILC, both generally and specifically concerning the current design of the Swedish SILC. In order to grasp all details concerning the LCS fully, the reader should also have some knowledge about the previous design of the LCS. We refer to (Eurostat, 2021) for general methodological

[1] The total sample size for living conditions statistics at Statistics Sweden is similar to before.

[2] To avoid time series breaks, we have used the same calibration procedure for SILC 2008 and onwards.

information on SILC and to (SCB, 2020) and (SCB, 2020) for information on the current design of the Swedish LCS and SILC.

## 1.1 Outline

In Section 2, we present the set of candidate variables for the new cross-sectional auxiliary vector. We describe the set of candidate variables with respect to categorization and origin. We also give some more detailed explanation of initial choices with respect to e.g., categorization, for some candidate variables.

In Section 3, we provide a brief description of our methodology with respect to general theory and the methods used. Section 4 describe the data used in the analyses.

In Sections 5–8, we present our analyses for the cross-sectional auxiliary vector, i.e., descriptive analysis, regression analysis, indicator analysis, and estimates of survey and register variables. Each section describes the methodology, and it is also briefly described in Section 3.

Section 9 describes how we derive auxiliary vectors for longitudinal estimation from the chosen cross-sectional auxiliary vector.

Finally, in Section 10, we briefly discuss some general issues pertaining to the present work.

# 2 Auxiliary variables

In this section, we introduce the set of candidate variables for the revised cross-sectional auxiliary vector. For all candidate variables, we present their current and possible categories, and we discuss some initial choices with respect to categorization and to other properties of our candidate variables

## 2.1 Candidate auxiliary variables

In Table 1, we show our candidate auxiliary variables. For each variable, we show the current SILC categorization (if applicable), current number of categories (for variables included in the current SILC auxiliary vector), possible other categorizations, register origin, and whether the current auxiliary vectors for SILC and LCS (with possibly different categorizations) include the variable. Candidate variables come from the total population register (TPR), the electoral register, which contains the electoral roll for Sweden, the education register (EDU), the longitudinal database for integration studies (STATIV), and the register of income and taxation (IoT). In total, there are twenty-two possible auxiliary variables, of which one or both current auxiliary vectors include thirteen. When we consider all possible categorizations of the variables, there are thirty candidate variables. Note that we do not consider the current LCS categorizations for candidate variables in Table 1.

## 2.2 Initial variable selection

The objective of the present evaluation is to select an auxiliary vector, which efficiently adjusts for non-response bias and reduces the variance of estimates. Consequently, we want to examine as many candidate variables as possible with respect to their performance when included in an auxiliary vector. However, already the initial set of candidate variables is subject to some variable selection, concerning, e.g., variable categorization. We outline the choice of categories and possible alternatives in Table 1. A candidate variable included in the current auxiliary vector will typically keep the same categorization as in the current vector. For some variables, we also propose alternative categorizations. For candidate variables not included in the current auxiliary vector, we suggest categories from important domains and previous experience on e.g., the expected size of categories. We provide some additional comments on the initial variable selection below.

We only consider candidate variables *Age* and *Sex* as the cross-classified variable *Age x sex*, because there are known differences between the sexes concerning e.g., response propensity, in particular for older age groups. In addition, this is a common cross-classification. Because of

differences in living conditions within ages 16–24 years, it is possible to split this group into ages 16–19 years and 20–24 years instead.

The candidate variable *Voter participation* refers to the latest parliament election. Newly arrived immigrants and younger person becoming eligible to take part in the survey between elections will not have been able to vote in this election. These groups will increase in size between elections; hence, the distribution of the variable will gradually change. In addition, the relevant information, i.e., if an individual voted or not, becomes less up to date as time elapses from the last election. At the same time, the information on newly arrived immigrants and younger persons, i.e., that they were not able to vote, is less relevant. Consequently, we choose to use three categories for this variable: qualified to vote and did vote; qualified to vote and did not vote; not qualified to vote.

The list of candidate variables includes an indicator of whether an individual has a disposable income above or below the median disposable income. We choose not to include an indicator of whether household disposable income is below or above the median household disposable income, since the candidate variable *Register-based AROP* utilizes household income in a similar fashion.

The candidate variable *Telephone number* is an indicator of whether a sample person has a known telephone number or not. Each sample person gets an indicator value when included in the sample for the first time. For most sample persons, the value stays the same throughout the survey, but it may change for persons for which all contact attempts failed, i.e., non-contacts. For *Telephone number*, we provide an alternative categorization in which we cross known or unknown telephone number with whether a sample person is included in the sample for the first time, i.e., is in the first wave, or not.

All candidate variables are on the individual level. Previously, we did not use auxiliary information on the household level; hence, consistency is one reason not to introduce such auxiliary information. In addition, register households differ from SILC households; in particular, some individuals do not belong to a register household. Furthermore, we will calibrate selected respondent personal weights separately from 2021 and onwards such that we may not use auxiliary information on the household level. Hence, the introduction of auxiliary information on the household level may result in inconsistencies between the weights of the survey.

**Table 1**
**Candidate auxiliary variables, current categorization in the SILC auxiliary vector (if applicable), current number of categories in the SILC auxiliary vector, possible categorizations, source register, and inclusion in current auxiliary vectors.**

| Candidate variable | Current SILC categories | Current no of categories | Possible categorizations | Register | SILC aux. variable | LCS aux. variable |
|---|---|---|---|---|---|---|
| NUTS2 regions | NUTS2 regions | 8 | - | TPR | Yes | No |
| Age x sex | Ages 0-5, 6-10, 11-15, 16-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and 85-years, all divided by sex | 22 | We may divide ages 16-24 into ages 16-19 and ages 20-24 because of differences in living conditions. | TPR | Yes | Yes |
| Voter participation | - | - | Qualified to vote and did vote, qualified to vote and did not vote, not qualified to vote. | Electoral register | No | No |
| Education level | Missing or not registered, basic education, secondary education, tertiary education, less than 16 years old | 5 | - | EDU | Yes | Yes |
| Civil status | Unmarried; married, registered partner; divorced, divorced partner; widow/widower, surviving partner; less than 16 years old | 5 | - | TPR | Yes | Yes |
| Telephone number | - | - | Known telephone number, not known telephone number; division by panel. | Sample | No | Yes |
| Foreign/Swedish Background | - | - | Foreign born, born in Sweden with two foreign born parents, born in Sweden with one foreign born parent and one Swedish born parent, born in Sweden with two Swedish born parents | TPR | No | No |
| Country of birth | Born in Sweden or not | 2 | Born in Sweden, born in other Nordic country, born in EU outside of the Nordic countries, born in other European country, born in non-European country | TPR | Yes | Yes |
| Duration of stay in Sweden | - | - | 0-9 years, 10+ years, born in Sweden; 0-3 years, 4-6 years, 7-9 years, 10+ years, born in Sweden | STATIV | No | No |
| Degree of urbanization | - | - | Cities, towns and suburbs, rural areas | TPR | No | No |
| Individual/household disposable income | Individual income deciles | 10 | Household income deciles | IoT | Yes | Yes |
| Individual/household disposable income | Disposable income value | - | Logarithm of disposable income value | IoT | Yes | No |
| Financial aid | Receives, does not receive | 2 | - | IoT | Yes | No |
| Sickness compensation | Receives, does not receive | 2 | - | IoT | Yes | No |
| Housing allowance | Receives, does not receive | 2 | - | IoT | Yes | No |
| Pension | - | - | Receives, does not receive | IoT | No | No |
| Student aid | - | - | Receives, does not receive | IoT | No | No |
| Register-based AROP | - | - | Indicator variable whether income is below or above the ROP60 threshold; division by NUTS2 | IoT | No | No |
| Median individual disposable income | - | - | Above or below median individual disposable income; division by NUTS2 | IoT | No | No |
| Household type | - | - | Single-person household, cohabiting household, other household | TPR | No | No |
| Household size | - | - | 1 person, 2 persons, 3-4 persons, 5+ persons | TPR | No | No |
| Number of children in household | - | - | No Children, 1-3 Children, 4+ Children | TPR | No | No |

# 3 Methodology

## 3.1 Selection of auxiliary vectors

In Sections 5–8, we will use different quantitative methods to evaluate the efficiency of candidate auxiliary vectors in order to find an auxiliary vector that efficiently corrects for non-response bias and reduces the variance of estimates. Informally, the criteria for an efficient auxiliary vector are (Särndal & Lundström, 2005):

1) The auxiliary vector should explain the response propensity
2) The auxiliary vector should explain the main study variables
3) The auxiliary vector should identify the most important domains.

We will rely on all three criteria during the initial selection of candidate vectors. We then evaluate the efficiency of these vectors for estimates of survey variables and estimates of register variables. Note that Table 1 represents variable selection with respect to criterion 3) such that our quantitative evaluation will primarily focus on criteria 1) and 2).

## 3.2 Quantitative methods

We will use a standard toolbox of quantitative methods to evaluate candidate variables and candidate vectors in a subsequent manner. In the initial analyses, we focus on the properties of individual candidate variables in order to get an overview of their performance as auxiliary variables. The result from the initial analyses serves as an input to the following analyses, in which we study candidate vectors.

In Sections 5 and 6, we focus on the individual candidate variables. We describe the properties of the candidate variables in Table 1 using descriptive methods and logistic regression. In the descriptive analysis, we evaluate candidate variables with respect to the expected number of respondents in variable categories, and we look at the estimated response propensity in the categories of candidate variables. In the logistic regression analysis, we evaluate candidate variables with respect to model fit in both simple and multiple regression models. We also evaluate possible interactions between candidate variables.

In Section 7, we shift our focus to candidate vectors and use indicators of non-response adjustment to select candidate vectors for further evaluation. We primarily use a stochastic stepwise selection method utilizing non-response indicators to select candidate variables to the candidate vectors. In Section 8.2, we evaluate these candidate vectors with respect to the distribution of calibrated weights. Our final evaluation in Sections 8.4 and 8.5 concern the efficiency of candidate vectors with respect to the properties of estimates of survey variables

and estimates of register variable. We evaluate estimates of survey variables with respect to variance and estimates of register variables with respect to bias, variance, and MSE. We select an auxiliary vector for the Swedish SILC based on the results from estimates of survey variables and estimates of register variables.

The evaluation of longitudinal auxiliary vectors in Section 9 sets out from the results of the evaluation of the cross-sectional auxiliary vector. We primarily evaluate candidate vectors with respect to the distributions of calibrated weights for the two-, three-, and four-year longitudinal weights. For the four-year weights, we also evaluate candidate vectors with respect to estimates *of persistent at-risk-of-poverty*, *PAROP*.

# 4 Data

In our evaluation of candidate variables, we will study the efficiency of variables with respect to response propensity and estimation of a sample of survey variables from SILC and LCS. We select the sample of survey variables together with subject matter experts. Because of the current design, these survey variables will come from separate databases, which contain survey variables included in both SILC and LCS, survey variables included in LCS only, and survey variables included in SILC only. We will merge data from multiple databases to construct "samples" with the new sample size of 20 000 sample persons.

In order to evaluate the relationship between candidate variables and response propensity, we will use selected respondents from SILC and LCS from 2016 to 2019. Each year, there are between 19 470 and 19 773 selected respondents. We use responding selected respondents and household members from the same period and sources to evaluate the relationship between candidate variables and survey variables included in both SILC and LCS. Each year, there are between 23 567 and 25 812 selected respondents and household members. Note that the candidate variable *Voter participation* is available only for 2019; the datasets from 2016 to 2018 will not contain *Voter participation*.

To evaluate the relationship between candidate variables and survey variables from the LCS, we use the joint LCS response set from two subsequent years. That is, we use the response set from 2016 and 2017, and from 2018 and 2019. The joint sets contain 11 635 and 11 149 respondents, respectively. Note that, because of question module rotation, neither joint response set contains all survey variables. Also, note that because the auxiliary variable *Voter participation* is available in 2019 only, it is not included in these data. Neither do they contain the auxiliary variable *Telephone number*, divided by panel, since the LCS does not use a panel design.

To evaluate the relationship between candidate variables and SILC survey variables, we use responding selected respondents and household members. Data contain the joint response sets of 2016 and 2017, the joint response sets of 2017 and 2018, and the joint response sets of 2018 and 2019, respectively. Each set contains between 27 873 and 29 022 selected respondents and household members. Note that, because the auxiliary variable *Voter participation* is available in 2019 only, it is not included. Data will also not contain the auxiliary variable *Telephone number*, divided by panel, since it is not obvious how to create it for joint response sets in SILC.

# 5 Descriptive analysis

In this section, we begin our quantitative analysis of candidate variables. That is, this is the first part of our evaluation of candidate variables and vectors with respect to criteria 1) and 2) of Subsection 3.1. In this section, we focus on the properties of individual candidate variables and not on candidate vectors.

## 5.1 Candidate variables and response propensity

In Table 2, we show the average number of respondents in the category with the least number of respondents and the average maximum difference between the estimated response rates of the categories of a candidate variable for all categorical candidate variables in Table 1 over the years 2016 to 2019. We have adjusted the number of respondents to a sample size of 20 010. To estimate the response rates, we used the design weight for selected respondents. For those variables where several categorizations are possible, we differ between the categorizations by their number of categories as described in Table 1. Note that the order of the candidate variables is different in Table 2 compared to Table 1.

**Table 2**
**Average number of respondents in the category with the least number of respondents, corrected for a sample size of 20 010, and average maximum difference of estimated response rates between variable categories.**

| Candidate variable | Average least no of respondents | Average maximum difference of response rates (p.p.) | Candidate variable | Average least no of respondents | Average maximum difference of response rates (p.p.) |
|---|---|---|---|---|---|
| Age x sex, 22 categories | 123 | 24.1 | Household income deciles | 927 | 16.4 |
| Age x sex 24 categories | 123 | 27.6 | Median individual income, 2 categories | 4 491 | 10.4 |
| Register-based AROP, 2 categories | 1 133 | 13.0 | Median individual income, 16 categories | 186 | 16.0 |
| Register-based AROP, 16 categories | 47 | 19.7 | NUTS2 regions | 391 | 4.6 |
| Foreign/Swedish Background | 283 | 13.1 | Pension | 3 489 | 11.8 |
| Housing allowance | 681 | 9.6 | Sickness compensation | 300 | 14.4 |
| Civil status | 587 | 14.8 | Student allowance | 979 | 4.1 |
| Country of birth, 2 categories | 1 703 | 9.0 | Telephone number, 2 categories | 586 | 17.5 |
| Country of birth, 5 categories | 219 | 15.2 | Telephone number, 4 categories | 188 | 18.8 |
| Financial aid | 231 | 9.2 | Duration of stay in Sweden, 3 categories | 616 | 9.3 |
| Number of children in household | 245 | 8.4 | Duration of stay in Sweden, 5 categories | 145 | 14.7 |
| Household size | 1 089 | 8.4 | Degree of urbanization | 2 270 | 0.7 |
| Household type | 1 023 | 10.5 | Education level | 125 | 29.5 |
| Individual income deciles | 696 | 27.0 | Voter participation | 565 | 20.2 |

In Table 2, we see that the candidate variable *Register-based AROP* with sixteen categories is the only variable for which the least number of respondents in a category is less than 100. Because of the small number of respondents in this category, we exclude *Register-based AROP* with sixteen categories from the forthcoming analyses.

The results in Table 2 indicate that several of the candidate variables have an association with response propensity. *Education level* has the largest difference between estimated response rates, 29.5 percentage points. For candidate variables *Age x sex*, *Individual income deciles* and *Voter participation*, the maximum difference in estimated response rates exceeds 20 percentage points. In Table 2, we also see that candidate variables *Degree of urbanization*, *Student allowance* and *NUTS2 regions* show small maximum differences between the estimated response rates.

## 5.2 Auxiliary variables and survey variables

In Table 3, we show nine dichotomous survey variables and their attribute of interest. The table also shows whether it is a survey variable in SILC, LCS, or both SILC and LCS. These survey variables make up a

sample of all survey variables, which represent the subject matter content of the survey.

**Table 3**
**Survey variables, their attribute of interest, and whether it is a survey variable in SILC, LCS, or SILC and LCS.**

| Survey variable | Attribute | Survey variable in SILC, LCS or SILC and LCS |
|---|---|---|
| AROPE | At risk of poverty or social exclusion | SILC |
| AROP | At risk of poverty | SILC and LCS |
| Cash margin | No capacity to face unexpected financial expenses | SILC and LCS |
| Manual labour | Physically demanding work | LCS |
| Employee/Self-employed | Employee/Self-employed working full-time or part-time | SILC and LCS |
| Good health | Good or very good health | SILC and LCS |
| Close friend | Have a close friend | LCS |
| Read or listened to books | Have been reading books every week for the past 12 months | LCS |
| Do not go out at night | Refrained from going out due to anxiety | LCS |

In Table 4, for each of the survey variables in Table 3, we show the maximum difference between the estimated proportions having the attribute of interest in the categories of a candidate variable for all categorical candidate variables. For survey variables included in both SILC and LCS and for survey variables solely included in SILC, Table 4 shows the average results based on several survey years. For survey variables that are solely included in LCS, the results come from one survey year. Note that estimates for *Telephone number* with four categories and *Voter participation* only pertain to survey variables that are included in both SILC and LCS; see Section 4.

**Table 4**
**The maximum difference between estimated proportions in variable categories for each survey variable.**

| | Maximum difference | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Candidate variable** | AROP (Average) | AROPE (Average) | Cash margin (Average) | Employee/Self-employed (Average) | Good health (Average) | Read or listened to books | Close friend | Manual labour | Do not go out at night |
| Age x sex, 22 categories | 28.8 | 31.4 | 19.8 | 85.6 | 47.0 | 40.0 | 37.2 | 35.6 | 27.0 |
| Age x sex 24 categories | 28.8 | 31.4 | 22.5 | 85.6 | 47.9 | 41.8 | 37.8 | 35.6 | 28.3 |
| Register-based AROP | 76.2 | 77.6 | 32.8 | 32.6 | 14.0 | 2.8 | 11.8 | 7.2 | 4.0 |
| Foreign/Swedish Background | 23.8 | 24.5 | 24.5 | 14.4 | 6.5 | 5.7 | 13.5 | 8.3 | 4.7 |
| Housing allowance | 37.6 | 42.2 | 38.9 | 40.2 | 26.8 | 2.8 | 12.8 | 12.8 | 10.7 |
| Civil status | 16.4 | 17.4 | 15.4 | 54.8 | 26.3 | 11.7 | 12.2 | 22.9 | 7.2 |
| Country of birth, 2 categories | 22.5 | 23.1 | 23.7 | 4.0 | 3.7 | 2.7 | 10.6 | 2.2 | 0.0 |
| Country of birth, 5 categories | 33.0 | 35.9 | 37.2 | 23.1 | 10.9 | 16.0 | 13.8 | 11.2 | 2.3 |
| Financial aid | 64.9 | 70.7 | 69.1 | 33.6 | 17.9 | 6.7 | 19.3 | 9.3 | 2.8 |
| Number of children in household | 30.9 | 30.1 | 25.1 | 28.9 | 11.8 | 11.0 | 6.9 | 10.1 | 0.3 |
| Household size | 14.7 | 16.8 | 16.0 | 31.1 | 16.8 | 11.6 | 6.2 | 10.2 | 4.0 |
| Household type | 17.8 | 20.6 | 20.5 | 14.8 | 11.1 | 5.7 | 1.3 | 6.3 | 4.3 |
| Individual income deciles | 44.3 | 46.9 | 33.0 | 66.2 | 28.7 | 12.4 | 12.9 | 29.0 | 9.9 |
| Household income deciles | 62.7 | 66.1 | 31.3 | 59.8 | 32.2 | 6.7 | 14.5 | 23.5 | 8.3 |
| Median individual income, 2 categories | 25.8 | 27.3 | 18.9 | 52.5 | 16.5 | 3.7 | 7.8 | 12.5 | 5.8 |
| Median individual income, 16 categories | 30.2 | 33.6 | 23.8 | 57.7 | 20.2 | 13.1 | 14.2 | 21.8 | 15.6 |
| NUTS2 regions | 7.4 | 8.4 | 4.5 | 11.6 | 5.6 | 10.4 | 4.7 | 6.6 | 7.5 |
| Pension | 2.6 | 2.7 | 9.0 | 59.5 | 18.5 | 10.8 | 10.9 | 22.9 | 0.4 |
| Sickness compensation | 13.9 | 22.2 | 28.2 | 29.4 | 48.9 | 2.0 | 14.3 | 10.4 | 10.3 |
| Student allowance | 13.5 | 16.3 | 9.7 | 36.7 | 10.7 | 7.9 | 6.5 | 11.0 | 5.5 |
| Telephone number, 2 categories | 40.9 | 44.0 | 29.1 | 27.9 | 9.1 | 3.9 | 9.9 | 10.0 | 1.6 |
| Telephone number, 4 categories | 44.2 | . | 30.2 | 31.6 | 9.9 | . | . | . | . |
| Duration of stay in Sweden, 3 categories | 40.6 | 41.4 | 37.4 | 9.1 | 11.7 | 2.7 | 16.1 | 5.0 | 6.7 |
| Duration of stay in Sweden, 5 categories | 54.9 | 56.6 | 45.5 | 28.5 | 16.4 | 7.9 | 19.4 | 9.6 | 8.4 |
| Degree of urbanization | 3.5 | 3.4 | 0.8 | 5.6 | 5.3 | 6.5 | 2.5 | 5.7 | 3.3 |
| Education level | 46.7 | 48.4 | 33.3 | 45.8 | 18.0 | 21.8 | 31.0 | 22.4 | 2.5 |
| Voter participation | 19.6 | . | 31.6 | 44.1 | 11.7 | . | . | . | . |

In Table 4, we see that several candidate variables show strong relationships with one or more survey variables. For the income-based survey variables *AROP*, *AROPE*, and *Cash margin*, income-based candidate variables, e.g., *Register-based AROP*, *Income deciles*, and *Financial aid*, show the strongest relationships in general. However, there are some differences between the survey variables, where, e.g., *Register-based AROP* shows a much stronger association with *AROP* and *AROPE* than with *Cash margin*. Candidate variables *Duration of stay in Sweden* and *Education level* also show strong relationships with the income-based survey variables.

The results in Table 4 show that *Age x sex* has a strong relationship with the survey variable *Employee/Self-employed*. *Age x sex* also has a strong relationship with the survey variable *Good health* and shows the largest maximum difference among candidate variables for all other survey variables, i.e. *Manual labour*, *Close friend*, *Read or listened to books*, and *Do not go out at night*. The candidate variable *Pension* shows a strong relationship with *Employee/Self-employed* and *Sickness compensation* shows a strong relationship with *Good health*. In general, the strength of the relationships between candidate variables and survey variables is weak for the "Quality of life" survey variables *Close friend*, *Read or listened to books*, and *Do not go out at night*.

Candidate variables *Degree of urbanization* and *NUTS2 regions* show weak associations with all survey variables.

# 6 Regression analysis

Logistic regression models are useful when evaluating candidate variables for an auxiliary vector. In addition to verifying the results from the descriptive analysis in Section 5, logistic regression models may provide further insight into the effects of including a candidate variable in the auxiliary vector. Logistic regression is also useful in evaluating possible cross-classifications of auxiliary variables for inclusion in candidate vectors. In the logistic regression analyses, we will evaluate the efficiency of candidate variables with respect to model fit statistics. We will use both response propensity and the survey variables in Table 3 as response variables in our models.

Throughout the logistic regression analysis, we will use likelihood ratio statistics to compare models including candidate variables with a model, which does not include any explanatory variables, or which includes a basic subset of candidate variables.

## 6.1 Simple logistic regression

We begin with simple logistic regression models, i.e., with only a single explanatory variable. In Table 5, we show the average likelihood ratio statistic[3] and the number of years for which a candidate variable was statistically significant in logistic regression models with response propensity as response variable. We also use *AROPE* for the years 2017-2019 and *Do not go out at night* from 2017 as response variables. We show results for all 28 remaining categorical candidate variables.

[3] The distribution of the average of likelihood ratio statistics is somewhat complex; therefore, we do not use it as a statistical measure, but rather as an indicative measure of variable importance.

**Table 5**
**Results from simple logistic regression analysis.**

| Response variable / | Response propensity | | | AROPE | | Do not go out at night | |
|---|---|---|---|---|---|---|---|
| Candidate variable | Degrees of freedom[4] | Average LR statistic | Significant years | Average LR statistic | Significant years | LR statistic 2017 | Significant years |
| Age x sex, 22 categories | 15 | 464 | 4 | 527 | 3 | 741 | 1 |
| Age x sex 24 categories | 17 | 478 | 4 | 539 | 3 | 743 | 1 |
| Register-based AROP | 1 | 168 | 4 | 9541 | 3 | 15 | 1 |
| Foreign/Swedish Background | 3 | 155 | 4 | 2053 | 3 | 9 | 1 |
| Housing allowance | 1 | 57 | 4 | 1137 | 3 | 67 | 1 |
| Civil status | 3 | 357 | 4 | 406 | 3 | 55 | 1 |
| Country of birth, 2 categories | 1 | 102 | 4 | 1629 | 3 | 0 | 0 |
| Country of birth, 5 categories | 4 | 130 | 4 | 2110 | 3 | 2 | 0 |
| Financial aid | 1 | 29 | 3 | 1187 | 3 | 1 | 0 |
| Number of children in household | 2 | 29 | 4 | 849 | 3 | 0 | 0 |
| Household size | 3 | 84 | 4 | 1019 | 3 | 26 | 1 |
| Household type | 2 | 159 | 4 | 1568 | 3 | 39 | 1 |
| Individual income deciles | 9 | 430 | 4 | 3734 | 3 | 121 | 1 |
| Household income deciles | 9 | 170 | 4 | 5119 | 3 | 58 | 1 |
| Median individual income, 2 categories | 1 | 213 | 4 | 2714 | 3 | 75 | 1 |
| Median individual income, 16 categories | 15 | 245 | 4 | 2760 | 3 | 126 | 1 |
| NUTS2 regions | 7 | 24 | 3 | 138 | 3 | 37 | 1 |
| Pension | 1 | 220 | 4 | 79 | 3 | 0 | 0 |
| Sickness compensation | 1 | 64 | 4 | 123 | 3 | 29 | 1 |
| Student allowance | 1 | 13 | 4 | 188 | 3 | 24 | 1 |
| Telephone number, 2 categories | 1 | 189 | 4 | 1505 | 3 | 1 | 0 |
| Telephone number, 4 categories | 3 | 205 | 4 | . | . | . | . |
| Duration of stay in Sweden, 3 categories | 2 | 108 | 4 | 2258 | 3 | 16 | 1 |
| Duration of stay in Sweden, 5 categories | 4 | 121 | 4 | 2411 | 3 | 18 | 1 |
| Degree of urbanization | 2 | 1 | 0 | 34 | 3 | 21 | 1 |
| Education level | 3 | 467 | 4 | 838 | 3 | 11 | 1 |
| Voter participation | 2 | 443 | 1 | . | . | . | . |

In Table 5, we see that the average likelihood ratio statistic differs between candidate variables for all response variables. We can also see that the magnitude of the average likelihood ratio statistic values differs between the response variables, where *AROPE* has the largest values and *Do not go out at night* has the smallest values. We can see that *Age x sex* has the highest value of the average LR statistic for response

[4] Because the current analysis uses selected respondents, not all categories may apply, and the degrees of freedom may be lower than the number of categories implies. E.g., we do not use age groups for ages 15 years and younger.

propensity and *Do no go out at night*, while *Register-based AROP* has the highest value for *AROPE*[5].

The results in the descriptive analysis and the logistic regression models show small differences between the two possible cross-classifications of Age x sex. In addition, the categorization where we divide ages 16-24 years into ages 16-19 years and 20-24 years do not produce smaller groups. Therefore, we will only use that categorization from now on.

### 6.1.1 Continuous income variables

From Table 1, we see that we may consider both individual and household disposable income as continuous variables. Because income variables have heavy-tailed distributions, continuous income variables may cause the model to fit badly; indeed, the Hosmer-Lemeshow statistic[6] show highly significant values ($X^2$ between 90 and 201 on 8 degrees of freedom) for both individual and household disposable income when used as a single continuous variable for all years. In addition, the Akaike information criterion (AIC) is lower for income deciles than continuous income for all years[7].

It is not clear how the diagnostic criteria for logistic regression models relate to calibration estimation. However, because income deciles are an important domain and, in many cases, constitute a powerful auxiliary variable, it is very likely to be included in the auxiliary vector. Moreover, because it is a categorical variable, income deciles are not subject to any problems with model fit; rather, such categorizations often reproduce non-linear relationships adequately. Consequently, we argue that the inclusion of income deciles in the auxiliary vector is sufficient, and do not consider continuous income variables in the following.

## 6.2 Multiple logistic regression

In multiple logistic regression, we include several candidate variables in the regression models, which makes it possible to study their joint effect in a candidate vector. The primary objective of the multiple logistic regression analysis is to identify possible interactions between variables. To identify useful interactions, we compare models, which include explanatory variables separately, with models, which include their interaction.

---

[5] The degrees of freedom of the averaged likelihood ratio statistics are different for different candidate variables and thus not directly comparable.

[6] The Hosmer-Lemeshow statistic tests whether the expected event rates in subgroups are equal to the observed event rates.

[7] We also considered a log transform of income coupled with an indicator variable of whether income was zero or negative. This is a common setup for modelling purposes and produced slightly better results than the pure income values.

We use an initial subset of candidate variables, which will be included in all models. Otherwise, the absence of other variables may make us end up falsely selecting interactions, which has little or no effect together with other variables, for further investigation. For our purposes, we include *NUTS2 regions*, *Age x sex*, *Education level*, and *Individual income deciles*[8] in the initial subset. Either these variables relate to the design of the survey, or they identify important domains, i.e., fulfil criterion 3) of Subsection 3.1.

### 6.2.1 Candidate vectors with one additional variable

We start by looking at models where we include the initial subset of variables and each of the other candidate variables separately. For every comparison, each response variable yields many results. Hence, we restrict our evaluation to the quality-of-life variables *Close friend* and *Do not go out at night*, for which it is more difficult to find strong relationships with individual candidate variables; see e.g. Table 3. We also evaluate candidate variables with respect to response propensity.

In Table 6, we show the results from the multiple logistic regression analyses for response variables response propensity, *Close friend*, and *Do not go out at night*. For response propensity, we see that, except for *Voter participation*, the values of the average likelihood ratio statistics are relatively low. For the survey variables, the values of the statistic are higher for *Close friend* than they are for *Do not go out at night*.

[8] One might argue that we should use *Household income deciles* instead of *Individual income deciles*, or that we should evaluate both in a similar fashion. However, while there is correlation between the two types of income, the correlation is far from one-to-one; for example, the Pearson correlation for the 2019 sample is 0.48. Because the main objective of this analysis is to identify important interactions, we feel that the current setup is adequate also for interactions including household income.

**Table 6**
**Results from multiple logistic regression with an initial subset of variables and an additional variable.**

| Response variable / | Response propensity | | | Close friend | | Do not go out at night | |
|---|---|---|---|---|---|---|---|
| Candidate variable | Degrees of freedom | Average LR statistic | Significant years | LR statistic 2017 | p-value 2017 (percent) | LR statistic 2017 | p-value 2017 (percent) |
| Register-based AROP | 1 | 5 | 2 | 25 | 0.0 | 2 | 12.9 |
| Foreign/Swedish Background | 3 | 27 | 4 | 36 | 0.0 | 12 | 15.6 |
| Housing allowance | 1 | 12 | 4 | 96 | 0.0 | 4 | 27.1 |
| Civil status | 3 | 85 | 4 | 15 | 0.0 | 25 | 0.0 |
| Country of birth, 2 categories | 1 | 15 | 3 | 40 | 0.0 | 23 | 0.0 |
| Country of birth, 5 categories | 4 | 45 | 4 | 96 | 0.0 | 0 | 69.3 |
| Financial aid | 1 | 12 | 2 | 140 | 0.0 | 2 | 81.8 |
| Number of children in household | 2 | 5 | 1 | 35 | 0.0 | 1 | 33.1 |
| Household size | 3 | 32 | 4 | 22 | 0.0 | 4 | 14.0 |
| Household type | 2 | 62 | 4 | 41 | 0.0 | 21 | 0.0 |
| Household income deciles | 9 | 39 | 4 | 27 | 0.0 | 22 | 0.0 |
| Median individual income, 16 categories | 7 | 5 | 0 | 11 | 30.8 | 23 | 0.7 |
| Pension | 1 | 3 | 1 | 8 | 32.3 | 11 | 13.8 |
| Sickness compensation | 1 | 11 | 4 | 0 | 58.6 | 0 | 92.6 |
| Student allowance | 1 | 23 | 4 | 28 | 0.0 | 18 | 0.0 |
| Telephone number, 2 categories | 1 | 54 | 4 | 0 | 76.2 | 0 | 82.0 |
| Telephone number, 4 categories | 3 | 72 | 4 | 13 | 0.0 | 0 | 56.6 |
| Duration of stay in Sweden, 3 categories | 2 | 62 | 4 | 132 | 0.0 | 23 | 0.0 |
| Duration of stay in Sweden, 5 categories | 4 | 92 | 4 | 139 | 0.0 | 26 | 0.0 |
| Degree of urbanization | 2 | 1 | 0 | 1 | 51.6 | 20 | 0.0 |
| Voter participation | 2 | 239 | 1 | . | . | . | . |

### 6.2.2 Candidate vectors with interactions between candidate variables

To evaluate interaction effects, we construct models with one pairwise interaction and the initial subset of variables as explanatory variables. We do this for all possible pairwise interactions. We do not consider interactions where the number of individuals in the smallest category is 50 or less. Because of the unknown effects of loss of relevance between elections, we do not consider interactions involving *Voter participation*. Neither do we consider interactions involving the constructed variable *Telephone number*. We restrict our evaluation to two-way interactions. Note that we always include the initial subset of variables in the model, even if a variable from the initial subset is a part of the interaction term. Otherwise, the likelihood ratio statistic might take negative values.

We see that no interaction terms show clear improvements to the model fit for neither response propensity, *Close friend* nor *Do not go out at night*. For some interactions, we see an increase in the likelihood ratio statistic compared to the models with only one additional variable in Subsection 6.2.1. However, this improvement is typically small, not significant or the result of an increased degrees of freedom. For the

remaining analyses, we consider no additional interaction terms except those already introduced in the previous sections.

## 6.3 Variable selection

The primary objective of the descriptive analysis in Section 5 and the regression analysis in Subsections 6.1 and 6.2 is to provide an overview of the efficiency of candidate variables. Variable selection is of less importance at this stage. We did however exclude a few candidate variables from the initial set of candidate variables in Table 1 from further analysis based on the results in Sections 5 and 0. We excluded the candidate variable *Register-based AROP, subdivided within NUTS2 regions*, because it has categories with too few individuals; see Subsection 5.1. From Subsection 6.1, we have that we use the categorization of *Age x sex* where we divide ages 16-24 into ages 16-19 years and ages 20-24 years because there is little difference compared to the non-subdivided variable. Because of the model fit results in Subsection 6.1.1, we excluded continuous income variables in the forthcoming analyses.

In addition, we will exclude *Degree of urbanization* from here and onwards because it shows very little effect on both response propensity and survey variables. After the selection of variables based on the results in Sections 5 and 6, twenty-four candidate variables remain.

# 7 Indicator analysis

In Sections 5 and 6, our focus was on individual candidate variables, and in some cases, cross-classifications of them. However, from this section and onwards, we focus on candidate vectors, i.e., subsets of the candidate variables in Table 1 excluding those variables described in Subsection 6.3. In this section, we use indicators of the efficiency of candidate vectors together with the results from the previous sections to select a subset of possible candidate vectors for further analysis.

## 7.1 Indicators

In this section, we use the $H_1$ and $H_3$ indicators (Särndal & Lundström, 2010). Both indicators serve as proxy values for the bias reduction obtained with respect to response propensity ($H_3$) and survey variables ($H_1$). Because response propensity affects all survey variables, the $H_3$ indicator may identify a general-purpose auxiliary vector, efficient for all or most survey variables. Briefly, the $H_3$ indicator corresponds to the coefficient of variation for the adjustment of the design weights due to the calibration procedure. A large $H_3$ value indicates a large adjustment and a corresponding bias reduction pertaining to all survey variables. The $H_1$ indicator shows the bias reduction of a specific survey variable. A large $H_1$ value indicates a large bias reduction for the survey variable.

In a stepwise approach, the $H_3$ value will gradually increase; however, the rate of the increase will typically decline substantially with the addition of more variables. The $H_1$ values will typically increase at first and then decline but may show an irregular behaviour. In our stepwise analyses, we will select a candidate vector using the $H_3$ indicator such that we add variables to the vector sequentially until the increase in the indicator value for the vector is less than a given tolerance value, which we set to 1 %. The candidate variables selected so far make up the chosen candidate vector. For the $H_1$ indicator, we select the candidate vector as the subset of variables selected in the stepwise procedure up until the indicator values start to decrease.

### 7.1.1 Variable selection with indicators

Indicator values provide an indication of the ability of a candidate vector to reduce bias. However, because we utilize several indicators, and because the candidate vector with the largest indicator values may not provide the best auxiliary vector with respect to estimation performance, i.e., mean squared error, we use the results from indicator analysis to select several candidate vectors. We evaluate the performance of the selected vectors with respect to the uncertainty of actual estimates in Section 8.

When selecting candidate variables for an auxiliary vector, stepwise approaches are common. However, stepwise approaches utilize a linear

optimization strategy and a greedy algorithm, and hence might possibly fail. We therefore propose a simple stochastic stepwise algorithm to aid in the selection of candidate vectors. In this algorithm, we randomly select $m = 10$ of our $p$ candidate variables and then perform stepwise selection as described in Subsection 7.1 from the $m$ selected variables. From many runs, we may derive the proportion of selected candidate vectors including a candidate variable and the average increase of indicator values due to the inclusion of a candidate variable. Note that if $m = p$, the algorithm is equivalent to the standard stepwise approach.

## 7.2 Results

In Table 7, we show the rate of inclusion in selected candidate vectors from the stochastic stepwise algorithm for the $H_3$ indicator for all candidate variables. Note that we do not show the results for *AROP* because the results for *AROPE* are very similar. We also show the corresponding results for all survey variables and the $H_1$ indicator. We calculate the inclusion rate as the proportion of candidate vectors including a candidate variable among those cases when it is one of the $m$ initially selected candidate variables. For variables where data come from several years, we show the average value over all years. Each year's values come from 100 runs of the stochastic stepwise algorithm.

In Table 8, we show the average increase of indicator values when a candidate variable is included in the chosen candidate vector for all candidate variables. The indicator values come from the same runs and setup of the stochastic stepwise algorithm as described previously for the results in Table 7.

In the tables, we can see that candidate variables *Age x sex*, *Individual income deciles*, *Education level*, and *Voter participation* have the highest inclusion rates and the highest average increase of indicator values for response propensity, i.e., the $H_3$ indicator. In addition, candidate variables *Telephone number* with two categories and *Duration of stay in Sweden* with five categories show large inclusion rates and average increases of indicator values.

**Table 7**
**Inclusion rate in candidate vectors selected from the stochastic stepwise algorithm using indicators for all candidate variables.**

| Response variable (indicator) / Candidate variable | Response propensity ($H_3$, average) | AROPE ($H_1$, average) | Cash margin ($H_1$, average) | Employee / Self-employed ($H_1$, average) | Good health ($H_1$, average) | Read or listened to books ($H_1$) | Close friend ($H_1$) | Manual labour ($H_1$) | Do not go out at night ($H_1$) |
|---|---|---|---|---|---|---|---|---|---|
| Age x sex, 24 categories | 100 | 100 | 98 | 12 | 36 | 100 | 58 | 100 | 8 |
| Register-based AROP | 13 | 100 | 74 | 3 | 4 | 2 | 28 | 2 | 8 |
| Foreign/Swedish Background | 39 | 84 | 79 | 0 | 4 | 14 | 2 | 18 | 13 |
| Housing allowance | 0 | 67 | 64 | 8 | 7 | 9 | 2 | 2 | 67 |
| Civil status | 88 | 51 | 40 | 0 | 0 | 67 | 42 | 26 | 100 |
| Country of birth, 2 categories | 0 | 38 | 47 | 0 | 4 | 0 | 0 | 0 | 0 |
| Country of birth, 5 categories | 48 | 34 | 29 | 0 | 5 | 18 | 10 | 4 | 0 |
| Financial aid | 10 | 9 | 34 | 5 | 6 | 10 | 10 | 2 | 0 |
| Number of children in household | 7 | 31 | 26 | 0 | 0 | 7 | 2 | 0 | 17 |
| Household size | 13 | 43 | 46 | 0 | 0 | 14 | 14 | 4 | 44 |
| Household type | 43 | 90 | 94 | 4 | 9 | 12 | 8 | 0 | 48 |
| Individual income deciles | 98 | 100 | 99 | 80 | 14 | 67 | 0 | 7 | 100 |
| Household income deciles | 47 | 66 | 71 | 6 | 28 | 22 | 6 | 11 | 33 |
| Median individual income, 2 categories | 42 | 56 | 50 | 27 | 44 | 29 | 10 | 38 | 54 |
| Median individual income, 16 categories | 65 | 58 | 57 | 25 | 39 | 42 | 11 | 10 | 52 |
| NUTS2 regions | 5 | 12 | 21 | 5 | 4 | 0 | 0 | 5 | 0 |
| Pension | 50 | 41 | 62 | 73 | 25 | 49 | 28 | 55 | 84 |
| Sickness compensation | 25 | 66 | 68 | 6 | 8 | 3 | 2 | 0 | 63 |
| Student allowance | 36 | 8 | 26 | 4 | 5 | 25 | 9 | 41 | 0 |
| Telephone number, 2 categories | 80 | 99 | 97 | 9 | 7 | 2 | 14 | 16 | 0 |
| Telephone number, 4 categories | 84 | - | 85 | 11 | 9 | - | - | - | - |
| Duration of stay in Sweden, 3 categories | 60 | 7 | 23 | 0 | 4 | 64 | 4 | 4 | 0 |
| Duration of stay in Sweden, 5 categories | 81 | 0 | 15 | 5 | 12 | 82 | 0 | 4 | 0 |
| Education level | 100 | 35 | 99 | 12 | 73 | 100 | 100 | 63 | 13 |
| Voter participation | 100 | - | 100 | 0 | 14 | - | - | - | - |

From the analysis based on the $H_1$ indicator values, we see that the candidate variable *Register-based AROP* shows a large inclusion rate and average increase of indicator values for the survey variable *AROPE*. In addition, candidate variables *Foreign/Swedish background* and *Household type* also show large values for *AROPE*. For the remaining survey variables, candidate variables *Civil status* and *Pension* emerge as strong candidate variables for several survey variables.

**Table 8**
**Average increase of indicator values in selected candidate vectors from the stochastic stepwise algorithm using indicators for all candidate variables.**

| Response variable (indicator) / Candidate variable | Response propensity ($H_3$, average) | AROPE ($H_1$, average) | Cash margin ($H_1$, average) | Employee / Self-employed ($H_1$, average) | Good health ($H_1$, average) | Read or listened to books ($H_1$, average) | Close friend ($H_1$, average) | Manual labour ($H_1$, average) | Do not go out at night ($H_1$, average) |
|---|---|---|---|---|---|---|---|---|---|
| Age x sex, 24 categories | 131 | 21 | 21 | 45 | 20 | 21 | 14 | 28 | 2 |
| Register-based AROP | 17 | 75 | 24 | 8 | 5 | 1 | 8 | 3 | 2 |
| Foreign/Swedish Background | 12 | 5 | 8 | 0 | 1 | 1 | 1 | 2 | 1 |
| Housing allowance | 0 | 3 | 5 | 1 | 2 | 0 | 0 | 2 | 3 |
| Civil status | 39 | 5 | 6 | 0 | 0 | 8 | 10 | 6 | 4 |
| Country of birth, 2 categories | 0 | 3 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Country of birth, 5 categories | 12 | 4 | 7 | 0 | 1 | 1 | 6 | 4 | 0 |
| Financial aid | 5 | 2 | 3 | 1 | 1 | 0 | 1 | 0 | 0 |
| Number of children in household | 7 | 2 | 3 | 0 | 0 | 1 | 0 | 0 | 1 |
| Household size | 7 | 7 | 5 | 0 | 0 | 1 | 2 | 5 | 2 |
| Household type | 10 | 10 | 10 | 1 | 3 | 0 | 2 | 0 | 3 |
| Individual income deciles | 78 | 44 | 27 | 63 | 17 | 6 | 0 | 11 | 11 |
| Household income deciles | 19 | 21 | 13 | 13 | 17 | 2 | 6 | 4 | 5 |
| Median individual income, 2 categories | 25 | 16 | 11 | 54 | 20 | 4 | 8 | 16 | 8 |
| Median individual income, 16 categories | 28 | 14 | 10 | 52 | 15 | 4 | 5 | 15 | 8 |
| NUTS2 regions | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| Pension | 36 | 8 | 12 | 60 | 18 | 10 | 10 | 25 | 3 |
| Sickness compensation | 6 | 3 | 3 | 2 | 9 | 0 | 2 | 0 | 2 |
| Student allowance | 10 | 3 | 1 | 5 | 2 | 2 | 1 | 5 | 0 |
| Telephone number, 2 categories | 24 | 6 | 8 | 4 | 2 | 1 | 6 | 4 | 0 |
| Telephone number, 4 categories | 24 | | 7 | 2 | 2 | | | | |
| Duration of stay in Sweden, 3 categories | 9 | 2 | 4 | 0 | 2 | 2 | 6 | 0 | 0 |
| Duration of stay in Sweden, 5 categories | 13 | 0 | 5 | 2 | 3 | 3 | 0 | 0 | 0 |
| Education level | 110 | 8 | 14 | 30 | 21 | 26 | 19 | 19 | 2 |
| Voter participation | 177 | | 35 | 0 | 5 | | | | |

### 7.2.1 Selection of candidate vectors

From our analysis of the indicator values from the stochastic stepwise algorithm described in Table 7 and Table 8, and from standard stepwise selection, individual indicator values, and the previous results in Sections 5 and 6, we suggest candidate vectors for further analyses.

Because candidate variables selected by the $H_3$ indicator are general-purpose and work for all survey variables, we suggest including strong

candidate variables identified in the $H_3$ analysis, i.e., *Age x sex*, *Individual income deciles*, *Education level*, and *Voter participation*, in all candidate vectors. In addition, we suggest including *NUTS2 regions* or *Median individual income*, 16 categories in all candidate vectors, since both candidate variables identify the stratification variable NUTS2 regions. This subset of candidate variables also forms a candidate vector for use in the coming analyses.

*AROPE* is the main European indicator on poverty and social exclusion, and an important goal of the design of the Swedish SILC is to fulfil the precision requirements for *AROPE*; in particular, the stratification and sample allocation is a consequence of this. Hence, we also suggest including *Register-based AROP* together with the subset of strong candidate variables from the $H_3$ analysis in a candidate vector.

Besides the already mentioned candidate variables, *Civil status* and *Pension* emerge as the two most important candidate variables. Therefore, we select a candidate vector as the subset of candidate variables from the $H_3$ analysis, *Register-based AROP*, *Civil status*, and *Pension*.

Among the rest of the candidate variables, we keep *Sickness compensation* and *Student allowance*. The remaining candidate variables, i.e., *Housing allowance*, *Country of birth*, *Financial aid*, *Number of children in household*, *Household size*, *Household income deciles*, and *Median individual income* show either no relation to response propensity or survey variables, or relate to them similarly as other, stronger, candidate variables. We will therefore not consider them further.

As our fourth candidate vector we hence choose subset of candidate variables from the $H_3$ analysis, *Register-based AROP*, *Civil status*, *Pension*, *Sickness compensation*, and *Student allowance*. In Table 9, we show the four candidate vectors. Logistic regression analyses with response propensity as response variable showed no apparent signs of e.g., collinearity issues for any of the candidate vectors.

**Table 9**
**Candidate vectors selected for further evaluation.**

| Candidate vector 1 | Candidate vector 2 | Candidate vector 3 | Candidate vector 4 |
|---|---|---|---|
| Age x sex, 24 categories | Age x sex, 24 categories | Age x sex, 24 categories | Age x sex, 24 categories |
| Individual income deciles | Individual income deciles | Individual income deciles | Individual income deciles |
| Education level | Education level | Education level | Education level |
| Voter participation | Voter participation | Voter participation | Voter participation |
| NUTS2 / NUTS2 x median | NUTS2 / NUTS2 x median | NUTS2 / NUTS2 x median | NUTS2 / NUTS2 x median |
| | Register-based AROP | Register-based AROP | Register-based AROP |
| | | Civil status | Civil status |
| | | Pension | Pension |
| | | | Telephone number, 2 categories |
| | | | Duration of stay in Sweden, 5 categories |
| | | | Foreign/Swedish Background |
| | | | Household type |
| | | | Sickness compensation |
| | | | Student allowance |

Note that we chose to include the candidate variable *Telephone number* with two categories in Candidate vector 4 because it is easier to construct and shows very similar performance as *Telephone number* with four categories in the analyses. Also, note that, because individual income variables performs better than household income variables and because *Register-based AROP* uses household income such that it is implicit in candidate vectors 2-4, we chose to include *Individual income deciles* only in the forthcoming analyses.

# 8 Estimates of survey and register variables

In Section 7.2.1, we present four candidate vectors based on the results from the descriptive analysis, logistic regression analysis, and indicator analysis. In this section, we will look at the performance of these candidate vectors with respect to estimates of survey variables and estimates of register variables. For survey variables, we consider the variance of estimates, and for register variables, we consider the bias, variance, and MSE of estimates. We will choose an auxiliary vector for the survey from the four candidate vectors based on these results.

## 8.1 Estimator

In the previous design of the Swedish SILC, each panel sample constituted a stratified network sample with a proportional sample allocation, where stratification was with respect to age of selected respondent in eight categories. In the cross-sectional estimation procedure, we considered all panels as a single stratified sample. The stratification variable was the same as for the separate panel samples. Because of the similarity of the samples, and because all samples were representative of the reference population, it is likely that this estimation procedure resulted in unbiased estimates despite its approximate nature.

From 2021 and onwards, NUTS2 region is the stratification variable for new panel samples. The sample allocation is non-proportional and may vary slightly between samples. Because the survey has a rotating panel design with six panels, panel samples with the previous sample design will be in the cross-sectional sample until 2025. When we have different sample designs in the cross-sectional sample, we cannot use the previous estimation procedure since it relies on the similarity of panel samples. Note that this also applies after 2025, because the sample allocation may vary between panels.

We suggest using a composite estimator for an estimated total $\hat{t}_y$, i.e.

$$\hat{t}_y = \sum_j c_j \hat{t}_{y_j},$$

where summation is over panel samples, $c_j$ is a constant, and $\hat{t}_{y_j}$ is the estimate from panel sample $j$. We choose the constant $c_j$ as the proportion of selected respondents in the cross-sectional sample that come from panel $j$. The composite estimator accounts for the different sampling designs and accurately reflects the rotating panel design.

## 8.2 Analysis of calibrated weights

We begin our analysis in this section by evaluating the properties of the calibration estimation weights using each of the four candidate vectors shown in Table 9. We analyze the variance of the weights as well as the occurrence of negative weights and unduly large weights. From now on, we exclude Telephone number from the set of candidate variables; see Section 8.3.

If the variance of the weights is large, it may indicate that estimates are subject to an increased variance due to calibration. This is typically the result of a too large auxiliary vector, i.e., with too many auxiliary variables and too many categories of the auxiliary variables. We compare variances between candidate vectors. Our analysis shows only a small increase of the variance of the weights with increased vector size and hence we conclude that our candidate vectors are similar with respect to the variance of the weights.

Negative weights or unduly large weights are undesirable, and one tries to ensure that an auxiliary vector does not result in such weights (Särndal & Lundström, 2005). Table 10 shows that when we do calibration estimation on the merged LCS and SILC sample from 2019, we get fourteen negative weights with vector 1, thirteen negative weights with vector 2, eighteen negative weights with vector 3 and thirty-two negative weights with vector 4.

Further analysis showed that the candidate variable *Voter participation*, and in particular, individuals in category *Not qualified to vote*, caused negative weights. To address this, we reduced the number of categories of *Voter participation* from three to two, which decreased the number of negative weights as shown in Table 10. In addition, our analysis showed that the candidate variable *Duration of stay in Sweden* caused negative weights. By removing this variable, the number of negative weights decreased from twenty-five to five for candidate vector 4.

**Table 10.**
**The number of negative weights generated when using different auxiliary vectors. The results come from the merged LCS and SILC sample from 2019.**

| | Negative weights |
|---|---|
| Vector 1 | 14 |
| Vector 2 | 13 |
| Vector 3 | 18 |
| Vector 4 | 32 |
| Vector 1 (Voter participation containing two categories) | 1 |
| Vector 2 (Voter participation containing two categories) | 1 |
| Vector 3 (Voter participation containing two categories) | 4 |
| Vector 4 (Voter participation containing two categories) | 25 |
| Vector 4 (Voter participation containing two categories and excluding Duration of stay in Sweden) | 5 |

## 8.3 Variable selection

Up until Section 8, we included the candidate variable *Telephone number* in the set of possible auxiliary variables. We also include it in Candidate vector 4. However, we did not fully consider all upcoming changes to the survey when we chose to include *Telephone number* in the present evaluation. In the present estimation procedure for the LCS, we construct response homogeneity groups from whether a sample person has a known telephone number or not. Because the survey will switch from CATI to mixed-mode CATI/CAWI in the survey year 2022, this is not a relevant approach from then and onwards. Consequently, we will not consider *Telephone number* further in this work; rather, we suggest considering other options for auxiliary information pertaining to data collection for the implementation of mixed-mode.

Th candidate variable *Median individual disposable income*, divided by NUTS2 did not cause negative weights and we therefore choose to include it in candidate vectors, and hence to exclude the candidate variable NUTS2 from here and onwards. In Table 11, we show our candidate vectors after these changes.

Note that it is not desirable to perform variable selection at this stage. To avoid this scenario, we could have calculated weights during the evaluation of candidate vectors in Section 7 to discover e.g., negative weights. We recommend doing so in future evaluations.

**Table 11.**
**Updated table of candidate vectors for further evaluation.**

| Candidate vector 1 | Candidate vector 2 | Candidate vector 3 | Candidate vector 4 |
|---|---|---|---|
| Age x sex, 24 categories | Age x sex, 24 categories | Age x sex, 24 categories | Age x sex, 24 categories |
| Individual income deciles | Individual income deciles | Individual income deciles | Individual income deciles |
| Education level | Education level | Education level | Education level |
| Voter participation | Voter participation | Voter participation | Voter participation |
| NUTS2 x median | NUTS2 x median | NUTS2 x median | NUTS2 x median |
| | Register-based AROP | Register-based AROP | Register-based AROP |
| | | Civil status | Civil status |
| | | Pension | Pension |
| | | | Foreign/Swedish Background |
| | | | Household type |
| | | | Sickness compensation |
| | | | Student allowance |

## 8.4 Estimates of survey variables

### 8.4.1 Survey variables

In this section, we evaluate our candidate vectors with respect to the efficiency of estimates of the nine survey variables described in Table 3. For estimates of survey variables, we only consider variance in our evaluation. We evaluate the estimated variance for almost 100 national and European domains.

Note that the survey variables *AROPE*, *Manual labour*, *Close friend*, *Read or listened to books* and *Do not go out at night* are applicable in either LCS or SILC. This means that we will not be able to include the auxiliary variable *Voter participation* in candidate vectors when estimating these survey variables. Also, note that we will use the composite estimator described in Section 8.1 when estimating survey variables included in both LCS and SILC.

### 8.4.2 Results

For each survey variable and candidate vector, we get almost 100 estimates. We count the number of times a candidate vector produces the lowest variance, the second lowest variance, and so on. We show the results in Table 12.

**Table 12.**
**The number of times each of the four vector gives the smallest variance, second smallest variance, second largest variance and largest variance.**

| Rank | Vector 1 | Vector 2 | Vector 3 | Vector 4 |
|---|---|---|---|---|
| Smallest variance | 323 (38 %) | 267 (31 %) | 69 (8 %) | 202 (23 %) |
| Second smallest variance | 205 (24 %) | 377 (44 %) | 226 (26 %) | 53 (6 %) |
| Second largest variance | 133 (15 %) | 167 (19 %) | 450 (52 %) | 111 (13 %) |
| Largest variance | 200 (23 %) | 50 (6 %) | 116 (13 %) | 495 (57 %) |

Table 12 shows that Vector 1 gives the smallest variance for most estimates; however, it is not obvious which vector that overall performs best. To test whether a vector generally gives smaller or larger variances compared to another vector, we have performed one-sided Wilcoxon rank-sum tests. We show the results in Table 13. The results in Table 13 indicate that vector 2 generally gives smaller variances than the other vectors. However, additional analysis shows that the difference is small.

**Table 13.**
**Results from one-sided Wilcoxon rank-sum tests.**

| Test ($H_1$) | Wilcoxon Statistic | p-value |
|---|---|---|
| Vector 1 gives larger variances than Vector 2 | $7.7*10^5$ | 0,00 |
| Vector 1 gives larger variances than Vector 3 | $6.5*10^5$ | 0,00 |
| Vector 1 gives larger variances than Vector 4 | $6.1*10^5$ | 0,00 |
| Vector 2 gives smaller variances than Vector 3 | $5.8*10^5$ | 0,00 |
| Vector 2 gives smaller variances than Vector 4 | $5.7*10^5$ | 0,00 |
| Vector 3 gives smaller variances than Vector 4 | $6.4*10^5$ | 0,00 |

Because we do not know the true value of survey variables, it is of limited value to compare point estimates. We show point estimates for SILC 2019 produced using the four candidate vectors and the published estimates in Table 14.

**Table 14**
**Point estimates (percent) for some survey variables using candidate vectors and published estimates from SILC 2019.**

| Survey variable | Vector 1 | Vector 2 | Vector 3 | Vector 4 | SILC 2019 |
|---|---|---|---|---|---|
| AROP | 16.6 | 16.5 | 16.6 | 16.5 | 17.1 |
| Cash margin | 20.7 | 20.6 | 20.7 | 20.4 | 20.0 |
| Employee / Self-employed | 56.7 | 56.8 | 56.8 | 57.0 | 56.6 |
| Good health | 76.0 | 76.0 | 75.9 | 75.7 | 74.6 |

## 8.5 Estimates of register variables

### 8.5.1 Register variables

In Section 8.4 we only considered the variance of estimates. In this section, we consider both the variance and bias[9] of estimates, since we know the true value of the estimated variables. In Table 15, we show the register variables, constructed dichotomous variables for use in the evaluation, and their register origin. We will produce the estimated bias and variance of estimates for the same domains as in Section 8.4.

**Table 15**
**Register variables used in the evaluation.**

| Register variable | Constructed variable | Register |
|---|---|---|
| Registered at the Swedish Public Employment Service | Registered at the Swedish Public Employment Service | STATIV |
| Employed | Employed | IoT |
| Family type | Cohabitating family | TPR |
| Country of birth | Born outside of the Nordic countries | TPR |
| Household size | Household size of three persons or more | TPR |
| Degree of urbanization | Living in Cities | TPR |

When estimating register variables, we use the merged LCS and SILC sample from 2019, since this sample is most like the sample that we expect to obtain in the new design. In addition, if we use the sample from 2019, we can include all auxiliary variables in the estimation and use the composite estimator described in Section 8.1.

### 8.5.2 Results

As in Section 8.4.2, we use aggregate measures, where we count the number of times a candidate vector gives the lowest variance, the second lowest variance, and so on. We use the same aggregate measure for the bias and the Mean Square Error (MSE). We show the results in Table 16. We see that, in general, candidate vectors 2 and 4 perform better with respect to variance, and candidate vectors 3 and 4 perform better with respect to bias. For example, for variance estimates estimated with candidate vector 2, 392 estimates had the lowest or

[9] The absolute value of the bias is used.

second lowest variance. In addition, most bias estimates from candidate vectors 3 and 4 are the lowest or the second lowest, and the opposite holds for candidate vectors 1 and 2.

For the MSE, 52 percent of the estimates with the lowest MSE come from candidate vector 4. In the Table, we also see that vector 4 gives the highest value of the MSE for 182 estimates. Further analysis shows that the increase in MSE in these estimates compared to the vector that gave the lowest MSE is small.

**Table 16.**
**The number of times each of the four vectors gives lowest variance, bias, and MSE, second lowest variance, bias, and MSE, second highest variance bias, and MSE and highest variance bias, and MSE.**

| Rank | Vector 1 | Vector 2 | Vector 3 | Vector 4 |
|---|---|---|---|---|
| Lowest variance | 50 (9 %) | 191 (35 %) | 13 (2 %) | 289 (53 %) |
| Second lowest variance | 195 (36 %) | 201 (37 %) | 131 (24 %) | 16 (3 %) |
| Second highest variance | 127 (23 %) | 140 (26 %) | 175 (32 %) | 101 (19 %) |
| Highest variance | 171 (31 %) | 11 (2 %) | 224 (41 %) | 137 (25 %) |
| Lowest bias | 81 (15 %) | 70 (13 %) | 122 (22 %) | 270 (50 %) |
| Second lowest bias | 102 (19 %) | 165 (30 %) | 225 (41 %) | 51 (9 %) |
| Second highest bias | 146 (27 %) | 247 (45 %) | 116 (21 %) | 34 (6 %) |
| Highest bias | 214 (39 %) | 61 (11 %) | 80 (15 %) | 188 (35 %) |
| Lowest MSE | 74 (14 %) | 84 (15 %) | 102 (19 %) | 283 (52 %) |
| Second lowest MSE | 116 (21 %) | 155 (29 %) | 229 (42 %) | 43 (8 %) |
| Second highest MSE | 135 (25 %) | 247 (45 %) | 126 (23 %) | 35 (6 %) |
| Highest MSE | 218 (40 %) | 57 (10 %) | 86 (16 %) | 182 (34 %) |

Table 17 shows one-sided Wilcoxon rank-sum tests comparing the estimated variance, bias, and from two candidate vectors. The results in Table 17 indicate that candidate vector 2 and candidate vector 4 generally gives lower variances than vector 1 and vector 3. The test does not give any strong indication that vector 2 gives lower variances than vector 4. The results in Table 17 also indicate that vector 4 generally gives lower bias and MSE than the other vectors.

**Table 17.**
**One-sided Wilcoxon rank-sum tests of hypotheses that estimates produced with one vector has higher bias/variance/MSE than estimates produced using another vector.**

| Test ($H_1$) | Wilcoxon Statistic | p-value |
|---|---|---|
| Vector 1 gives higher variances than Vector 2 | 3.6 | 0,00 |
| Vector 1 gives lower variances than Vector 3 | 2.7 | 0,00 |
| Vector 1 gives higher variances than Vector 4 | 3.4 | 0,00 |
| Vector 2 gives lower variances than Vector 3 | 2.0 | 0,00 |
| Vector 2 gives lower variances than Vector 4 | 2.9 | 0,15 |
| Vector 3 gives higher variances than Vector 4 | 3.6 | 0,00 |
| Vector 1 gives higher bias than Vector 2 | 3.3 | 0,00 |
| Vector 1 gives higher bias than Vector 3 | 3.4 | 0,00 |
| Vector 1 gives higher bias than Vector 4 | 3.3 | 0,00 |
| Vector 2 gives higher bias than Vector 3 | 3.2 | 0,00 |
| Vector 2 gives higher bias than Vector 4 | 3.2 | 0,00 |
| Vector 3 gives higher bias than Vector 4 | 3.1 | 0,02 |
| Vector 1 gives higher MSE than Vector 2 | 3.3 | 0,00 |
| Vector 1 gives higher MSE than Vector 3 | 3.4 | 0,00 |
| Vector 1 gives higher MSE than Vector 4 | 3.4 | 0,00 |
| Vector 2 gives higher MSE than Vector 3 | 3.1 | 0,00 |
| Vector 2 gives higher MSE than Vector 4 | 3.2 | 0,00 |
| Vector 3 gives higher MSE than Vector 4 | 3.1 | 0,00 |

## 8.6 Selection of a cross-sectional auxiliary vector

The analyses in Sections 5-7 led to the selection of four candidate vectors. In Subsections 8.2-8.5, we compared these vectors with respect to the efficiency of estimates of survey variables and estimates of register variables. The results show that, in general, candidate vector 4 performs best in terms of variance, bias, and MSE. We therefore select candidate vector 4 as the cross-sectional auxiliary vector for the Swedish SILC.

# 9 Longitudinal estimation

The revision of the cross-sectional estimation scheme described in Sections 5-8 sets out from the previous auxiliary vector. The longitudinal estimation scheme of the Swedish SILC however currently does not utilize any auxiliary information aside from stratification variables and estimates of longitudinal survey variables are therefore Horvitz-Thompson estimates. To further reduce bias due to non-response and possibly reduce variance, it is desirable to introduce additional auxiliary information through calibration estimation for longitudinal estimates.

Because the new longitudinal estimation procedure uses calibration estimation, it will be like the cross-sectional estimation procedure in practice. We first define a longitudinal frame population as the intersection of the corresponding cross-sectional frame populations. From this population, we may calculate e.g., stratum sizes and auxiliary population totals. We create longitudinal samples in a similar fashion.

## 9.1 Evaluation of longitudinal candidate vectors

The evaluation of the cross-sectional auxiliary vector utilized custom data as described in Section 4. For the evaluation of the longitudinal estimation scheme, we will use the SILC 2019 longitudinal data, since it is difficult to create data, which resemble the longitudinal panel structure. In addition, between the years 2021-2026, the requisites for the longitudinal estimation will change each year due to an increasing number of panels, changing sample size, and new panel sampling design, which will require a yearly revision of the longitudinal estimation scheme. Hence, the present evaluation may serve as a starting point and may be instructive for the upcoming revisions.

In our evaluation of the longitudinal estimation scheme, we intend to set out from the results from the evaluation of the cross-sectional auxiliary vector in Sections 5-8. We intend to use the chosen auxiliary vector from Section 8 as an initial candidate vector for all longitudinal auxiliary vectors. Because the number of respondents will decrease from the cross-sectional response set to the two-year longitudinal response set, and from the two-year longitudinal response set to the three-year longitudinal response set and so on, we expect to reduce the initial auxiliary vector as our evaluation progresses. Because of this process, we will in general not compare several candidate vectors, and instead select the largest vector, which produce acceptable weights.

The main longitudinal indicator is *Persistent at-risk-of-poverty (PAROP)*, which we estimate for the four-year longitudinal population. We will hence use *PAROP* in the evaluation of the four-year longitudinal estimates. In addition, we will introduce a register-based *PAROP*,

similarly as for the register-based *AROP* in the cross-sectional auxiliary vector, in the four-year vectors. Because we at present produce no longitudinal estimates pertaining to the two-year population or the three-year population, we will only evaluate the properties of the longitudinal weights for them similarly as in Section 8.2. Note that, because we do not compare candidate vectors as in Section 8.2, we do not compare the variances of weights.

## 9.2 Results

### 9.2.1 Two-year longitudinal estimates

We begin by using the selected auxiliary vector from the cross-sectional evaluation as candidate vector. When using this vector, we get thirty negative weights. Initial analysis of the two-year longitudinal response set showed that the number of individuals in one or several categories of candidate variables *Age x sex*, *Education level*, and *Median individual disposable income*, divided by NUTS2, was small. We therefore reduced the number of categories of *Age x sex* to sixteen by considering ages 0-15 years, 16-24 years, and 75 years and older as one category each instead of several categories as before. We also reduced the number of categories of *Education level* to four by letting individuals with *Missing or not registered education* and *Basic education* make up one category. We also replaced the subdivided *NUTS2 regions* by the *NUTS2 regions*. This reduced the number of negative weights to thirteen.

Further analysis showed that in order to reduce the number of negative weights to a small number, e.g., less than ten, we should use candidate vector 2 from Table 11 as our auxiliary vector, where we modified the categorization of variables *Age x sex*, *Education level*, and *NUTS2 regions* as described previously.

### 9.2.2 Three-year longitudinal estimates

Because the three-year longitudinal response set is smaller than the two-year response set, we must reduce the auxiliary vector further compared to the chosen vector for the two-year weights in order to achieve a reasonable adjustment in our calibration procedure. Specifically, we reduced the number of categories of *Age x sex* to ten, letting ages 25-44 years, ages 45-65 years, and ages 65 years and older make up three age categories instead of six as in the two-year vector. In addition, we had to exclude *Voter participation*, *Individual income deciles*, and *Register-based AROP* from the auxiliary vector.

The three-year longitudinal auxiliary vector hence consists of the modified *Age x sex* variable, *Education level*, for which the categories are the same as for the two-year vector, *Household type*, and *NUTS2 regions*, with the same categories as in the two-year vector.

### 9.2.3 Four-year longitudinal estimates

As mentioned in Subsection 9.1, we will introduce a register-based *PAROP* in the four-year longitudinal candidate vectors. Similarly, as for the two-year and three-year auxiliary vectors, we however need to reduce the four-year vector further to get an acceptable number of negative weights in our estimation process. For the four-year vector, we reduce the number of age categories of *Age x sex*, such that the age categories become 0-15 years, 16-34 years, 35-64 years, and 64 years and older. The other candidate variables remain the same as for the three-year vector.

We will use *PAROP* in our evaluation of the four-year estimation. Because we chose a candidate vector by reducing the three-year longitudinal vector, and because we do not use any register variables in our evaluation, we do not evaluate the chosen vector by comparing with, e.g., smaller vectors as in the cross-sectional case. Because we now introduce calibration estimation for longitudinal estimates, it is however of interest to compare with previous estimates. In Table 18, we compare previous point estimates from SILC 2018 and SILC 2019 with point estimates from SILC 2019 using calibration estimation for some domains. We see in the table that there is an increase in the calibrated point estimates for PAROP compared to previous years for most domains. Note that the variation in PAROP estimates between years is large also for other years.

**Table 18**
**Point estimates (percent) for PAROP from SILC 2018, SILC 2019, and SILC 2019 with the new auxiliary vector.**

| Domain | SILC 2018 | SILC 2019 | SILC 2019, Calibration estimation |
|---|---|---|---|
| Total | 5.7 | 7.4 | 9.9 |
| Male | 5.2 | 7.5 | 10.3 |
| Female | 6.1 | 7.2 | 9.6 |
| 18-24 years old | 4.6 | 16.3 | 19.2 |
| 25-49 years old | 3.8 | 5.7 | 8.3 |
| 50-64 years old | 5.0 | 4.3 | 4.5 |
| 65 years and older | 8.6 | 8.8 | 12.2 |

# 10 Discussion

In this work, we present a new auxiliary vector for cross-sectional estimation in the Swedish SILC. We chose the new auxiliary vector from a set of candidate variables through sequential analysis using multiple methods. We also select auxiliary vectors for longitudinal estimation. The new auxiliary vectors come from evaluation using data from 2016 to 2019. Because we will use the new auxiliary vectors in production from 2021 and onwards, one may view the implementation of the new vectors as part of the evaluation, i.e., the final step of the evaluation.

As briefly described in Section 1, the Swedish SILC will have a new design from 2021. The new design includes an increased number of panels, an increased sample size, and a new panel sampling design. The new design also features further integration between SILC and the Swedish LCS. Since it is not possible to fully implement the changes in the number of panels and sample size in one year, there will be a transition period, in which we e.g., use samples which only take part in the survey for one year, and in which the old and new panel sampling design exist simultaneously in the cross-sectional sample. We expect to fully implement the new design from 2026 and onwards. We do however not expect the transition period to affect the introduction of a revised auxiliary vector or vice versa.

Another major change to the survey is the introduction of mixed-mode data collection with web and telephone interviews in 2022, which is likely to affect estimates. The transition to mixed-mode is likely to include an experiment in 2022, which e.g. may reduce the size of the sample available for estimation, and which may motivate temporary changes to the auxiliary vector. In order to adjust for the effects of mixed-mode data collection on estimates, it is possible that we need to revise the auxiliary vector after the introduction of mixed-mode. The new data collection mode may also provide opportunities to create new auxiliary variables pertaining to e.g., preferred mode choice.

# References


European Union, 2019. *REGULATION (EU) 2019/1700 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.* [Online]
Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R1700&from=EN

Eurostat, 2021. *METHODOLOGICAL GUIDELINES AND DESCRIPTION OF EU-SILC TARGET VARIABLES, 2021 operation,* Luxembourg: European Commission.

SCB, 2018. *Utredningen om ny design för SILC/ULF,* Örebro: SCB.

SCB, 2020. *Kvalitetsdeklaration, Undersökningarna av levnadsförhållanden.* [Online]
Available at: https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/#_Dokumentation

SCB, 2020. *Resultatrapport estimationsprojektet,* Örebro: SCB.

SCB, 2020. *Statistikens framställning, Undersökningarna av barns levnadsförhållanden.* [Online]
Available at: https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/#_Dokumentation

Särndal, C.-E. & Lundström, S., 2005. *Estimation in Surveys with Nonresponse.* Chichester: John Wiley & sons.

Särndal, C.-E. & Lundström, S., 2010. Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology,* 36(2), pp. 131-144.

**Statistics Sweden describes Sweden**
Statistics Sweden provides society with statistics for decision-making, debate and research, on behalf of the Government, government agencies, researchers and industry. These statistics contribute to fact-based public discourse and informed decisions.