PHF User Guide Version 4.0

Updated: Mai 2023

Authors: PHF Survey Team

Table of contents

1 Introduction

The German Panel on Household Finances (PHF) is a triennial panel survey on household finances and wealth in Germany, covering the balance sheet, pensions, income, work life and other demographic characteristics of private households living in Germany. Until 2023 data have been collected for four survey waves, these were carried out in 2010/2011, 2014, 2017, and 2021 respectively.

Anonymised PHF data of the four waves are available as scientific use files (SUF) for the purpose of conducting scientific research. These can be provided to researchers upon direct request. The PHF scientific use files currently available are SUF Wave 1 Version 4.0, SUF Wave 2 Version 5.0, SUF Wave 3 Version 3.0, and SUF Wave 4 Version 1.0. In addition, two interim surveys with a smaller sample size were conducted in 2019 and 2020, between the main surveys, mainly for the purpose of panel maintenance. The corresponding scientific use files, SUF Interim Survey 2019 Version 1.0 and SUF Interim Survey 2020 Version 1.0, are also available for the research community. This documentation is a reference point for data users and is meant to facilitate their work with the PHF data. It first gives an overview of the structure and content of the available scientific use files. Then, it provides and discusses a variety of helpful insights into the specifics of the survey and the PHF data preparation steps. **It is highly recommended that all data users read this documentation.**

The remainder of the document is organised as follows:

The next two sections provide an overview of the structure of the most recent SUF and discuss the changes between the current and the previous version of the SUF. Section 4 describes the structure of the questionnaire and provides information about the main changes to the questionnaire across waves.

Section 5 provides a general overview of the data preparation steps undertaken by the Deutsche Bundesbank in order to improve the quality of the survey data for empirical analysis. This encompasses data editing as well as the multiple imputation of missing values. To meet the legal requirement of producing a *de-facto* anonymised data set for academic research, several anonymization steps have been performed. These are also laid out in section 5.

Section 6 focuses on the most important identifiers for households and individuals as well as on the coding of variables. Moreover, section 6 details the variables in the SUF related to the sampling design of the survey.

All changes to the original data can be tracked in the flag variables. They are introduced in detail in section 7. Section 8 focuses on weights and discusses calibrated weights as well as replicate weights for estimation. The final section provides a basic summary of the interim surveys, including their purpose, how they differ from the main survey, and the type of content they focus on.

The topics covered in this document are far from exhaustive and may be amended, supplemented and expanded over time. Hence, data users are invited to let us know whether

there is any important information missing that is relevant for their work with the PHF data and should be included in this user guide.

2 Data files of the scientific use file

2.1 H-file

The h-file contains variables that have been collected at the household level by interviewing the financially knowledgeable person (FKP).

2.2 P-file

The p-file contains variables that have been collected at the individual level by interviewing all household members aged 16 or older.

Proxy interviews were allowed, i.e. it was possible that another household member answered the personal level questions for an absent household member; the variables **pe9020**, **pf9020** and **pg9020** contain information as to whether this was the case for any of the three personal level sections.

Some household members aged 16 or older could not be interviewed during the field phase in person and no information was collected via proxy interviews (wave 1: 476 household members, wave 2: 423 household members, wave 3: 545 household members, wave 4: 910 household members). The information in the p-file, i.e. sections 7 (employment), 8 (pensions and insurance policies) and 9 (income), are completely imputed for these persons. The variable "**missing**" identifies those persons and has the value "1", if no personal interview was conducted.

2.3 M-file

The m-file contains basic demographic information about all household members, including children under the age of 16 years.

The household matrix variables (**vsmq***) indicate relationships between household members: for example, the variable **vsmq01** contains the relationship of household members to the person with **pid**=1 (available if **pid**>1), **vsmq02** contains the relationship to the person with **pid**=2 (available if **pid**>2) etc. (For a definition of the variable **pid**, see section 6.1.2.).

2.4 W-file

The w-file contains replicate weights (at the household level) for variance estimation. For a detailed description of the replicate weights, see section 8.3.

2.5 D-file

The d-file contains aggregate, or "derived" variables, which are computed based on the variables from the questionnaire. They mainly comprise important variables that are frequently used in the analysis of household finances (e.g. total assets, total outstanding balance of household liabilities, net wealth). The respective variables are derived for each implicate (for definition and information regarding implicates see section 5.2 about multiple

imputation) separately. See separate document “Panel on Household Finances (PHF) - List of derived variables” for the construction of the variables.

3 Available PHF scientific use files

All PHF scientific use files (SUF) are registered at DataCite and have a unique Digital Object Identifier (DOI, an international standard for data citing). You can use the DOI provided at DataCite to cite the specific PHF data set you worked with in your publications.

- Weblink to DataCite: http://www.datacite.org/
- Alternatively, you can find the DOI at the German website da-ra of the GESIS Leibniz-Institut für Sozialwissenschaften and the ZBW–Leibniz-Informationszentrum Wirtschaft: http://www.da-ra.de/de/fuer-forscher/daten-recherchieren/

3.1 Wave 1

3.1.1 SUF Wave 1 Version 4.0

The PHF scientific use file Wave 1 Version 4.0 data set released in 11/20/2019 is an updated version of the SUF Wave 1 Version 3.0 data set. It consists of five Stata files (see also section 2 above):

1. PHF_h_wave1_v4_0.dta: household level file
2. PHF_p_wave1_v4_0.dta: individual (16 and older) level file
3. PHF_m_wave1_v4_0.dta: household matrix and information on children <16
4. PHF_d_wave1_v4_0.dta: derived variables
5. PHF_w_wave1_v4_0.dta: replicate weights file for variance estimation

The corresponding DOI of the PHF scientific use file Wave 1 version 4.0 data set is “10.12757/Bbk.PHF.01.04.01”.

Changes in SUF Wave 1 Version 4.0 compared to the previous version SUF Wave 1 Version 3.0 (see below) are as follows:

- **anzhhm16** (number of household members aged 16 or older): For one household, the value of **anzhhm16** is not in line with the age structure of the household. That might occur, if the age **(ra0300)** of household members was edited or imputed. For that household, **anzhhm16** was adjusted.
- As far as possible, the definition of the derived variables were harmonised and standardised across waves. As a result, the definition of the derived variables in the PHF dataset in general accord with the definition of derived variables in the Household Finance and Consumption Survey (HFCS) dataset. However, in some cases differences could occur. For more details, see the documentation on the derived variables (Panel on Household Finances (PHF) - List of derived variables).

3.1.2 Previous Version: SUF Wave 1 Version 3.0

The PHF scientific use file Wave 1 Version 3.0 data set is an updated version of the SUF Wave 1 Version 2.0 data set and was released in 05/29/2019. It consists of five Stata files (see also section 2 above):

1. PHF_h_wave1_v3_0.dta: household level file
2. PHF_p_wave1_v3_0.dta: individual (16 and older) level file
3. PHF_m_wave1_v3_0.dta: household matrix and information on children <16
4. PHF_d_wave1_v3_0.dta: derived variables
5. PHF_w_wave1_v3_0.dta: replicate weights file for variance estimation

The corresponding DOI of the PHF scientific use file Wave 1 version 3.0 data set is "10.12757/PHF.01.03.01.stata".

Changes in SUF Wave 1 Version 3.0 compared to the previous version SUF Wave 1 Version 2.0 (see below) are as follows:

Household variables (h-file)

- **dhd510$x** ("Year of formation"):
  Items **dhd5102** as well as **dhd5103** in the loop for the three companies with the highest value accidentally contain the same values as **dhd5101**. This error was corrected.

Derived variables (d-file)

- **da2108** ("Other financial assets"):
  In cases when households answer items **dhd2310** (other securities in securities account) and **dhd0910** (market value of certificates in total) but not item **hd1920** (total value other assets), variable **da2108** reports a wrong value. For these households **da2108** was accidentally set to the sum of **hd1920** and **dhd0910** instead of the sum of **dhd0910** and **dhd2310**. This coding error was corrected. The variables **da2100** (Total financial assets 1, excl. public and occupational pension plans), **da3001** (Total assets 1, excl. public and occupational pension plans), and **dn3001** (Net wealth) also contain information about other financial assets (**da2108**). Erroneous values for these variables were corrected too.
- **dl2200** ("Payments for non-collaterised debt (flow)"):
  The variables **dhb1000** (leasing instalment for leased cars on which the household makes the payments - amount), and **dhc0110** (payments for other leasing contracts - amount) were accidentally not added computing variable **dl2200**. This error was corrected. The variable **dl2000** was also updated since it is computed by using **dl2200**.

3.1.3 Previous Version: SUF Wave 1 Version 2.0

The PHF scientific use file Wave 1 Version 2.0 data set is the first updated version of the wave 1 PHF data set and was released in 08/29/2017. It consists of the five Stata files (see also section 2 above):

1. PHF_h_wave1_v2_0.dta: household level file

2. PHF_p_wave1_v2_0.dta: individual (16 and older) level file,
3. PHF_m_wave1_v2_0.dta: household matrix and information on children <16
4. PHF_d_wave1_v2_0.dta: derived variables
5. PHF_w_wave1_v2_0.dta: replicate weights file for variance estimation

The corresponding DOI of the PHF scientific use file Wave 1 version 2.0 data set is "10.12757/ PHF.01.02.01.stata".

Changes in SUF Wave 1 Version 2.0 compared to the first release of SUF Wave 1 (see below) are as follows:

- The SUF 1 Version 2.0 contains an additional data file PHF_d_wave1_v2_0.dta with derived variables.
- The updated data set includes additional IDs (**wave, persid**).
- The coding of the four-digit-flags was adjusted to the coding of the four-digit-flags in wave 2 to get comparable flags for both waves.

Household variables (h-file)

- **dhb0120** ("Year household moved into main residence"): Further editing for people aged 90.

Household matrix (m-file)

- **vsmq16*** ("Relationships between the household members"):
  An additional category "siblings (adoptive siblings, step-siblings)" was included. Now, the number of categories is identical for wave 1 and wave 2.
- **ra0200** ("Gender") and **ra0300** ("Age"): some values were edited.

Variables on individuals (p-file)

- Additional filter checks and editing of imputed values of the following variables. In particular, the editing was carried out based on the following rules:
  1) **dpe1100** ("End date of last job"):
     Missing answer on the end date of the last job (**dpe1100**) should only be imputed if person has been employed before (**pe0900**=1). However, for some cases they were imputed in version 1, even if **pe0900** was not equal to one. This has been corrected. No information on the end date of a person's last job is provided if this person has never been employed.
  2) **dpf0710d/e** ("Letter detailing pension amount"):
     A missing answer on the amount of expected pension (**dpf1000d/e**) should only be imputed if the person had not received a letter on the pension amount (**dpf0710d/e** = -6). If respondents had not received a letter on the pension amount (**dpf0710d/e** = -6), they were asked about the amount of expected pension (**dpf1000d/e**). According to the structure of the questionnaire, **dpf1000d/e** should only be filled if **dpf0710d/e** = -6. However, **dpf0710d/e** has been imputed even in the case of **dpf0710d/e** = -6. Therefore, **dpf0710d/e** was set back to -6 if **dpf1000d/e** was filled and **dpf0710d/e** was imputed.

3.1.4 Previous Version: SUF Wave 1

The PHF scientific use file Wave 1 is the first version of the wave 1 PHF data set, which was released in March 2013. It consists of the four Stata files PHF_h.dta, PHF_p.dta, PHF_m.dta, and PHF_w.dta.

The corresponding DOI of the PHF scientific use file Wave 1 data set is "10.12757/PHF.01.01.01.stata".

3.2 Wave 2

3.2.1 SUF Wave 2 Version 5.0

The PHF scientific use file Wave 2 Version 5.0 data set released in 25/05/2023 is an updated version of the SUF Wave 2 Version 4.0 data set. It consists of the five Stata files:

1. PHF_h_wave2_v5_1.dta: household level file
2. PHF_p_wave2_v5_1.dta: individual (16 and older) level file,
3. PHF_m_wave2_v5_1.dta: household matrix and information on children <16
4. PHF_d_wave2_v5_1.dta: derived variables
5. PHF_w_wave2_v5_1.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 2 Version 5.0 data set is "10.12757/Bbk.PHF.02.05.01".

Changes in SUF Wave 2 Version 5.0 compared to SUF Wave 2 Version 4.0 are as follows:

- The variable **dra0550** ("In Germany in 2010") and its flag were added to the p- and m-file.

3.2.2 Previous Version SUF Wave 2 Version 4.0

The PHF scientific use file Wave 2 Version 4.0 data set released in 11/20/2019 is an updated version of the SUF Wave 2 Version 3.0 data set. It consists of the five Stata files:

1. PHF_h_wave2_v4_0.dta: household level file
2. PHF_p_wave2_v4_0.dta: individual (16 and older) level file,
3. PHF_m_wave2_v4_0.dta: household matrix and information on children <16
4. PHF_d_wave2_v4_0.dta: derived variables
5. PHF_w_wave2_v4_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 2 Version 4.0 data set is "10.12757/Bbk.PHF.02.04.01".

Changes in SUF Wave 2 Version 4.0 compared to SUF Wave 2 Version 3.0 are as follows:

- **anzhhm16** (number of household members aged 16 or older)
  For some few households, the value of **anzhhm16** is not in line with the age structure of the household. That might occur, if the age (**ra0300**) of household members was edited or imputed. In such cases, **anzhhm16** was adjusted.

- As far as possible, the definition of the derived variables were harmonised and standardised across waves. As a result, the definition of the derived variables in the PHF dataset in general accord with the definition of derived variables in the Household Finance and Consumption Survey (HFCS) dataset. However, in some cases differences could occur. For more details, see the documentation on the derived variables (Panel on Household Finances (PHF) - List of derived variables).

3.2.3 Previous Version: SUF Wave 2 Version 3.0

The PHF scientific use file Wave 2 Version 3.0 data set is an updated version of the SUF Wave 2 Version 2.0 data set and was released in 05/29/2019. It consists of the five Stata files:

1. PHF_h_wave2_v3_0.dta: household level file
2. PHF_p_wave2_v3_0.dta: individual (16 and older) level file,
3. PHF_m_wave2_v3_0.dta: household matrix and information on children <16
4. PHF_d_wave2_v3_0.dta: derived variables
5. PHF_w_wave2_v3_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 2 Version 3.0 data set is "10.12757/Bbk.PHF.02.03.01".

Changes in SUF Wave 2 Version 3.0 compared to SUF Wave 2 Version 2.0 are as follows:

Household variables (h-file)

- **kt_in_hh** ("Indicator for internal fkp"): **kt_in_hh** was accidently set to missing for split and refresher households. This error was corrected.
- **dhh510$xa-o** ("Donor of gift / inheritance"):
  The questionnaire first asks for the member of the household that received a gift or inheritance (item **dhh50$xa-o** recipient of gift / inheritance) and afterwards for the relationship between recipient and donor (**dhh510$xa-o**). According to structure of the questionnaire, the relationship eg for the recipient named in **dhh501a** should be stored in **dhh5101a,** for the recipient named in **dhh501b** should be stored in **dhh501a** etc. If the pid of the recipient changed between the first and the second wave, this was not given due to an editing error. This error was corrected

Derived variables (d-file)

- **da2108** ("Other financial assets"):
  In cases when households answer items **dhd2310** (other securities in securities account) and **dhd0910** (market value of certificates in total) but not item **hd1920** (total value other assets), variable **da2108** reports a wrong value. For these households **da2108** was accidentally set to the sum of **hd1920** and **dhd0910** instead of the sum of **dhd0910** and **dhd2310**. This coding error was corrected. The variables **da2100** (Total financial assets 1, excl. public and occupational pension plans), **da3001** (Total assets 1, excl. public and occupational pension plans), and **dn3001** (Net wealth) also contain information about other financial assets (**da2108**). Erroneous values for these variables were corrected too.

- **dl2200** (“Payments for non-collaterised debt (flow)”):
  The variables **dhb1000** (leasing instalment for leased cars on which the household makes the payments - amount), and **dhc0110** (payments for other leasing contracts - amount) were accidentally not added computing variable **dl2200**. This computation error was corrected. The variable **dl2000** was also updated since it is computed by using **dl2200**.

3.2.4 Previous Version: SUF Wave 2 Version 2.0

The PHF scientific use file Wave 2 Version 2.0 data set is the first updated version of the wave 2 PHF data set and was released in 08/29/2017. It consists of the five Stata files:

1. PHF_h_wave2_v2_0.dta: household level file
2. PHF_p_wave2_v2_0.dta: individual (16 and older) level file,
3. PHF_m_wave2_v2_0.dta: household matrix and information on children <16
4. PHF_d_wave2_v2_0.dta: derived variables
5. PHF_w_wave2_v2_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 2 Version 2.0 data set is “10.12757/Bbk.PHF.02.02.01”.

Changes in SUF Wave 2 Version 2.0 compared to SUF Wave 2 Version 1.0 are as follows:

- The imputation algorithm was improved (especially with regard to the gross/net income conversion). The SUF is based on an imputation using income variables processed by an updated tax conversion module (see section 3.3.2). To ensure consistency between the imputed income data and the rest of the imputed values, the complete imputation was re-done.

Household variables (h-file)

- The cross-sectional household weights **exhoch_hh** and **exw_hh** were newly calibrated using updated population statistics. Therefore, **exhoch_hh** and **exw_hh** slightly differ from SUF Wave 2 Version 1.0**.** Furthermore, the SUF now also contains longitudinal household weights (**wlong**).
- Variable **dhd0620** (“Saved sum for home loan savings - amount”) was accidentally set to zero for all observations. The new version contains the correct values.

Household matrix (m-file)

- For some people that already participated in the first wave of the PHF, variable **ra0500** (“How long have you been living in Germany”) accidentally contains the values from wave 1. The new version contains the correct values. Furthermore, code “-5” (Since birth, without long interruption) was replaced by the age of the interviewee.

Replicate weights (w-file)

- The w-file has been updated, ie the weights have been recalibrated so that their total equals the total of **exhoch_ww**.

3.2.5 Previous Version: SUF Wave 2 Version 1.0

The PHF scientific use file Wave 2 Version 1.0 is the first version of the wave 2 PHF data set which was released in May 2016. It consists of the four Stata files: PHF_h_wave2_v1_0.dta, PHF_p_wave2_v1_0.dta, PHF_m_wave2_v1_0.dta, and PHF_w_wave2_v1_0.dta.

The corresponding DOI string of the PHF scientific use file Wave 1 Version 1.0 data set is “10.12757/Bbk.PHF.02.01.01”.

3.3 Wave 3

3.3.1 SUF Wave 3 Version 3.0

The PHF scientific use file Wave 3 Version 3.0 data set released in 25/05/2023 is an updated version of the SUF Wave 3 Version 2.0 data set. It consists of the five Stata files:

1. PHF_h_wave3_v3_0.dta: household level file
2. PHF_p_wave3_v3_0.dta: individual (16 and older) level file,
3. PHF_m_wave3_v3_0.dta: household matrix and information on children <16
4. PHF_d_wave3_v3_0.dta: derived variables
5. PHF_w_wave3_v3_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 3 Version 3.0 data set is “10.12757/Bbk.PHF.03.03.01”.

Changes in SUF Wave 3 Version 3.0 compared to SUF Wave 3 Version 2.0 are as follows:

- The variable **dra0550** (“In Germany in 2010”) and its flag were added to the p- and m-file.
- Sampling variables **stratum** and **stich** were deleted from the p-file
- The labels of the variable **hg0800** for the values 2 "will rise about as much as the cost of living" and 3 "will rise less than the cost of living", which have been previously accidentally reversed, were corrected.

3.3.2 SUF Wave 3 Version 2.0

The PHF scientific use file Wave 3 Version 2.0 data set released in 11/20/2019 is an updated version of the SUF Wave 3 Version 1.0 data set. It consists of the five Stata files:

1. PHF_h_wave3_v2_0.dta: household level file
2. PHF_p_wave3_v2_0.dta: individual (16 and older) level file,
3. PHF_m_wave3_v2_0.dta: household matrix and information on children <16
4. PHF_d_wave3_v2_0.dta: derived variables
5. PHF_w_wave3_v2_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 3 Version 2.0 data set is “10.12757/Bbk.PHF.03.02.01”.

Changes in SUF Wave 3 Version 2.0 compared to SUF Wave 3 Version 1.0 are as follows:

- **anzhhm** (Number of household members) and **anzhhm16** (number of household members aged 16 or older)
  In some rare cases, the value of **anzhhm** is not in line with the members of the household in the survey. Furthermore, for some few households, the value of **anzhhm16** is not in line with the age structure of the household. That might occur, if the age (**ra0300**) of household members was edited or imputed. In such cases, **anzhhm** and **anzhhm16** were adjusted.
- **pfa100$x** (Expected year of payment – private and occupational pension plans)
  The variables from the loop on the expected year of payment of private and occupational pension plans were added to the data set.
- **dda2109** (Voluntary pension/whole life insurance)
  The variable **dda2109** should correspond to the value of a household's whole-life insurances, private pension plans and independently concluded contract. However, in some cases also contracts concluded by the employer were erroneously included in the computation of **dda2109**. This error was corrected.
- **dda2100** (Total financial assets 1 (excl. public and occupational pension plans))
  Variable **da2106 (Managed accounts)** was accidentally included twice in the computation of **da2100** whereas **da2108 (Other financial assets)** was accidentally not included in the computation of **da2100**. This error was corrected.
- As far as possible, the definition of the derived variables were harmonised and standardised across waves. As a result, the definition of the derived variables in the PHF dataset in general accord with the definition of derived variables in the Household Finance and Consumption Survey (HFCS) dataset. However, in some cases differences could occur. For more details, see the documentation on the derived variables (Panel on Household Finances (PHF) - List of derived variables).

3.3.3 Previous Version: SUF Wave 3 Version 1.0

The PHF scientific use file Wave 3 Version 1.0 data set is the first version of the wave 3 PHF data set and was released in 05/29/2019. It consists of the five Stata files:

1. PHF_h_wave3_v1_0.dta: household level file
2. PHF_p_wave3_v1_0.dta: individual (16 and older) level file,
3. PHF_m_wave3_v1_0.dta: household matrix and information on children <16
4. PHF_d_wave3_v1_0.dta: derived variables
5. PHF_w_wave3_v1_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 3 Version 1.0 data set is "10.12757/Bbk.PHF.03.01.01".

3.4 Wave 4

3.4.1 SUF Wave 4 Version 1.0

The PHF scientific use file Wave 4 Version 1.0 data set is the first version of the wave 4 PHF data set and was released in 25/05/2023. It consists of the five Stata files:

1. PHF_h_wave4_v1_0.dta: household level file
2. PHF_p_wave4_v1_0.dta: individual (16 and older) level file,
3. PHF_m_wave4_v1_0.dta: household matrix and information on children <16
4. PHF_d_wave4_v1_0.dta: derived variables
5. PHF_w_wave4_v1_0.dta: replicate weights file for variance estimation

The corresponding DOI string of the PHF scientific use file Wave 4 Version 1.0 data set is "10.12757/Bbk.PHF.04.01.01".

3.5 Interim Surveys

3.5.1 SUF Interim Survey 2019 Version 1.0

The PHF scientific use file Interim Survey 2019 Version 1.0 data set is the first version of the Interim Survey 2019 PHF data set and was released in 25/05/2023. It consists of the Stata file PHF_Interim_2019_v1_0.dta.
The corresponding DOI string of the PHF scientific use file Interim Survey 2019 Version 1.0 data set is "10.12757/Bbk.PHF.int.2019.01.01".

3.5.2 SUF Interim Survey 2020 Version 1.0

The PHF scientific use file Interim Survey 2020 Version 1.0 data set is the first version of the Interim Survey 2020 PHF data set and was released in 25/05/2023. It consists of the Stata file PHF_Interim_2020_v1_0.dta.
The corresponding DOI string of the PHF scientific use file Interim Survey 2020 Version 1.0 data set is "10.12757/Bbk.PHF.int.2020.01.01".

4 Questionnaire

Table 1 shows the main sections of the PHF questionnaire. Before the main survey starts, the household and interviewer identify (through a series of questions) the so-called "financial knowledgeable person" (FKP) in the household. This is the person who has the best overview of the household's finances. The FKP answers the questions in section 1 to 6 that refer to the household as a whole and cover topics such as socio-demographics, household consumption, balance sheet information, expectations and inheritances. In addition, the survey collects data on the employment and old age provision of each household member aged 16 years or older in section 7 and 8. Answers in section 9 on income may refer to the household as a whole or only to household members aged 16 or older, depending on the type of income reported.

Table 1: Modules of the PHF questionnaire

Section	Topic	Questions addressed to:
1	Socio-demographics	Financial knowledgeable person (FKP) → household level
2	Consumption	
3	Real estate and its financing	
4	Unsecured debts and financial constraints + beliefs, expectations, financial literacy	
5	Business wealth, liquid assets, financial assets	
6	Gifts and inheritances	
7	Employment	Individual household members 16+ → individual level
8	Pensions	
9	Income	FKP or individual household members 16+

The questionnaires are available on our PHF website:

In English:

- Wave 1: https://www.bundesbank.de/resource/blob/617388/c239a0ed2dca23c30e9d43026119aa3d/mL/phf-codebook-en-data.pdf
- Wave 2: https://www.bundesbank.de/resource/blob/617362/bdaaf1a0043ed9ba4ec26f92764299e1/mL/phf-codebook-wave2-en-data.pdf
- Wave 3: https://www.bundesbank.de/resource/blob/798120/a608fd1998cd4ecfafaa31b9771278d7/mL/phf-codebook-wave3-en-data.pdf
- Wave 4: https://www.bundesbank.de/resource/blob/892716/301702cf3d39d71ce9bc64bc3db56e64/mL/phf-codebook-wave4-en-data.pdf
- Interim Survey 2019 https://www.bundesbank.de/resource/blob/902296/0ed8e43eed4c4a271ff2332660ea091a/mL/phf-codebook-2019-en-data.pdf

- Interim Survey 2020
  https://www.bundesbank.de/resource/blob/829670/79b67e7a08062354bca6e8b5bad30789/mL/phf-codebook-2020-en-data.pdf

In German:

- Wave 1:
  https://www.bundesbank.de/resource/blob/617368/b0ee105da2e3919f38ddb69a1bd73da3/mL/phf-codebook-de-data.pdf
- Wave 2:
  https://www.bundesbank.de/resource/blob/617344/62a1001a29335589d411940282c7ce79/mL/phf-codebook-wave2-de-data.pdf
- Wave 3:
  https://www.bundesbank.de/resource/blob/798118/117e900b9583f5380e846cc7d1e42cff/mL/phf-codebook-wave3-de-data.pdf
- Wave 4:
  Add Questionnaire Link 4 here
- Interim Survey 2019:
  https://www.bundesbank.de/resource/blob/904872/c6f0496f1c1eb27cb08f448860c46fff/mL/phf-codebook-2019-de-data.pdf
- Interim Survey 2020:
  https://www.bundesbank.de/resource/blob/829900/27e996afd0dc3603f61228fecf62c4c2/mL/phf-codebook-2020-de-data.pdf

Table 2 provides a brief overview of the main additions to the fourth wave questionnaire in comparison with the third wave questionnaire.

Table 2: Most important differences in the questionnaire between wave 3 and wave 4

Section	New, significantly modified, and deleted questions
2	**New questions on:** • Make ends meet – Review **(dhi0850a-d)** • Financial losses - Coronavirus **(dhi1000a-d)** • Estimate losses wage income - Coronavirus **(dhi1010)** • Estimate losses other income - Coronavirus**(dhi1020)** • Estimate other financial losses - Coronavirus **(dhi1030)** • Level of savings **(dhi0410)**
3	New question on loan financing at time ownership was transferred **(dhb0850)** **Significantly modified questions:**

	• Questions on expectations – real estate prices – household main residence – renter with specific percentage categories **(dhb1370a-e, dhb1371a-e)** is replaced by non-percentage categories **(dhb1370)** and a percentage specification **(dhb1371)** • Questions on expectations – real estate prices – household main residence – owner with specific percentage categories **(dhb1350a-e, dhb1351a-e)** is replaced by non-percentage categories **(dhb1350)** and a percentage specification **(dhb1351)** • Question on third party support for property household main residence acquisition **(dhnb0100a-e)** is split into two variables: Support purchase main residence **(dhnb0101)** and Support purchase main residence - type **(dhnb0102a-d)** **Deleted questions:** • Mortgage credit – banking group **(dhb090$xa-g)**
4	**New questions on:** • Payments that are more than 90 days past due **(hc1270)** • Personal insolvency **(dhc1500)** • Reasons for not applying for loan/credit **(dhc1450)** • Expectations - Rent **(dhni1200)** • Expectations – Rent - Percent **(dhni1250)** • Propensity to save - Risk **(hiz0040b)** **Significantly modified questions:** • Question on late or still outstanding repayments for loans **(dhc0300)** is replaced by two questions: Late or missed loan repayments **(hc1251)** and Late or missed loan repayments - type **(hc1252)** • Question on price level expectations – Percentage **(dhni0855a-e)** with different answer categories is replaced by the same question with a numeric value in percent as an answer **(dhni0850)** **Deleted questions:** • Satisfaction with health **(zi102)** • Expectations for stock market – percentage **(dhni1150)**
5	New question on Online Banking **(dhnd0400)** **Deleted questions:**

	• Support from third parties in the formation of the company **(dhd520$xa-e)** • Support from the government in the formation of the company **(dhd560$x)** • Securities account – banking group **(dhd0900a-f)** • Cash **(dhd1400)** • Amount of cash **(dhd1410)** • Investment behaviour – interest rates **(dhd2970a-e)** • Crisis – realised gains / losses **(dhd1800)** • Crisis – concerns about investing in certain forms of assets **(hnd3040)** • Crisis – concerns about investing in certain forms of assets (text) **(dhd1900)** • Crisis – confidence in commercial banks **(dhnd0300)**
6	New question on Literacy – Diversification B **(dhnm0350)**
7	New question on Restrictions working life - Coronavirus **(dpe1210a-f)**
8	New question on duration of contribution payments – private pension provision **(dpf145$x)**

Table 3 provides a brief overview of the main additions to the third wave questionnaire in comparison with the second wave questionnaire.

Table 3: Most important differences in the questionnaire between wave 2 and wave 3

Section	New, significantly modified, and deleted questions
1	• New question on economic education **(dpa0450)**
2	**New questions on:** • Expenditures for travelling and trips **(hi0230)** • Propensity to save - unexpected lottery win **(hiz0040a)** • Estimate of household's net wealth position **(dhi0750)**
3	**New questions on:** • Year of construction – household main residence (HMR) **(dhb0105)** • Planned move-out – HMR **(dhb0125)** • Mortgage credit - bank group **(dhb090$xa-g)** • Other property mortgages - assignment **(dhb750$xa-f)**

	Significantly modified questions: • Question on HMR – expectations prices **(dhb1300)** is replaced by two questions: HMR - expectations prices – tenant **(dhb1370a-g)** and HMR - expectations prices – owner **(dhb1350a-g)** **Deleted Questions:** • Reasons for moving **(dhb0130a-o)** • Renegotiation of household main residence (HMR) mortgages **(hb115$x)** • Additional borrowing regarding HMR mortgages **(dhb1501)** • Financing of the planned property purchase **(dhb3100a-e)** • Reasons for being a renter **(dhb3200a-l)** • Renegotiation of mortgages for other properties **(hb315$x)** • Additional borrowing regarding other property mortgages **(hb3501)** • Number of other vehicles **(dhb1200a-h)**
4	• New question on satisfaction with health condition **(zi102)** **Deleted questions:** • Guarantees made **(dhc0400, dhc0410)** • Expectations change of taxes and social security contributions change over the next twelve months **(dhni0100)**
5	**New questions on:** • Agricultural or forestry enterprise **(intt2$x)** • Securities account - bank group **(dhd0900a-f)** • Planning horizon **(hnd4000)** **Significantly modified questions:** • Specific questions on certificates were deleted (see below). Questions on other securities in securities account include certificates. • Question on price level expectations - percentage **(dhni0855a-e)** replaces the previous question on price level expectations - percentage **(dhni0850)** **Deleted questions:** • Saving in the form of certificates **(dhd0775b, dhd0910, dhd1000, dhd1010, dhd1011)** • Crisis - realised gains / losses **(dhd1810)** • Crisis - change in net assets **(hnd3100)**

	• Crisis -consulting by commercial banks **(dhnd0400)** • Investment behaviour - selection of products **(dhd2950a-c)**
6	**New questions on:** • Recipient of contribution **(dhh5000a-o)** • Donor of contribution **(dhh5100a-o)** • Literacy - compound interest (ii) **(dhnm0400)**
7	**New questions on:** • Education father and mother **(dpe9200, dpe9210)** • Probability to lose job **(pez010)** • Probability to find a job **(pez020)** **Deleted questions:** • Crisis – type of expected changes or deterioration in job conditions **(pne2850a-f)** • Crisis - change and deterioration in job conditions **(pne2700)** • Crisis – type of change and deterioration in job conditions **(dpe1600a-f)**
8	**New questions on:** • Expected year of payment - statutory public old-age provision **(pfa1000d,e,m)** • Letter of pension level - occupational pension scheme - lump sum payment **(dpf072$x)** • Expected year of payment - private and occupational old-age provision **(pfa100$x)** • Modified question: expected total amount of payments from all pensions schemes - percentage **(dpf0900)** is now replaced by the question only including public and occupational pension schemes - percentage **(pfa1300)**
9	**New questions on:** • Income from statutory public pension - previous year **(dpg0550)** • Income from private or occupational pensions - previous year **(dpg0710)** • Income from private support **(hg0250)** • Amount of income from private support **(dhg0255)**

Table 4 provides a brief overview of the main additions to the second wave questionnaire in comparison with the first wave questionnaire.

Table 4: Most important differences in the questionnaire between wave 1 and wave 2

Section	New questions on
2	• Total non-durable consumption expenditures (**hi0220**) added
3	• Planned real estate acquisition (**dhb3000**) and financing (**dhb3100a-e**), • Reason for being a renter (**dhb3200a-l**) • Cost of telecommunication (**dhb0330, dhb0335**) • Ground rent – "Erbpacht" (**dhb0340, dhb0350, dhb0355**) • Additional information on household main residence (**dhb0410, dhb1300, dhb1400, dhb550$x**) • Purchase of vehicles (**dhb4800, dhb4810**)
4	• Expectations about real estate market prices, interest rates, taxes and stock markets (qualitative and quantitative) (**dhni0850, dhni0900, dhni0950, dhni1000, dhni1050, dhni1100, dhni1150**). • Additional information on loans from private individuals (**hc035$xa-l, dhc400$x, dhc410$x, hc036$x, dhc420$x, dhc430$x, dhc370$x, dhc371$x, dhc3900, dhc3910, dhc3800, hc0370**)
5	• Cash holdings (**dhd1400, dhd1410**), • Additional questions on investment behaviour (**dhd2950a-c, dhd2970a-e**). • Maturities of bonds (**dhd2515**)
7	• Country of birth and country of birth parents (**dpe9100, dpe9150**)
8	• Private and occupational old age provisions are collected for each contract, while in wave 1 information is only collected for classes of pension contracts • New questions on expected pension replacement rate (**dpf0900, dpf0950**).
9	• Principal income earner (**dhg2000a-b**)

5 Conversion of the Bundesbank's PHF data into a scientific use file

The process of converting the raw survey data collected by the interviewers into a scientific use file for academic research includes the following steps: editing of the data, imputation of missing values, conversion of variables, and anonymization of the data. This section describes the four steps in more detail below. As a result, there are some differences between the questionnaire and variables included in the scientific use file (SUF). Moreover, some variables in the questionnaire are not part of the SUF. This section mainly builds on information provided in Kalckreuth et al (2012).

5.1 Editing

The editing process aims to detect and remove inconsistencies, improve data quality, and prepare the data for imputation without distorting the data. In general, the process of editing the PHF can be divided into three steps: 1) all interviews were subjected to initial filter and value checks, i.e. mechanical data checks to ensure correct filtering of the answering path and the assignment of correct codes; 2) logical consistency checks, i.e. checks that tested the consistency of households' answers with other answers given during the interview; and 3) outlier checks that detected whether some values for a given household were clearly too high or low in comparison with the other answers given by the same household and with respect to other households in the data set. Edited values are marked by editing flags and are provided to users of the data (for further details on flag variables, see section 7).

All data in foreign currencies were converted into euro. Furthermore, some of the variables collected as verbatim responses were converted into numerical codes (e.g. NACE, ISCO) or were recoded into categorical answers by editors.

5.1.1 Occupation

The occupation is coded following the International Standard Classification of Occupation ISCO 88. The special code 79 means "manufacturing: supervisor / foreman". In most cases, two digits of ISCO are provided.

5.1.2 Industry of economic activity

The industry of economic activity is coded following NACE Rev. 2 from 2008. NACE categories A-U are recoded to numbers 1-21.

5.2 Multiple Imputation

In order to deal with item non-response, missing observations of all major PHF variables are imputed. You can identify imputed values by their corresponding flag variable having a value greater than or equal to 2000 (for more details on flag variables, see section 7).

The imputation procedure is based on the "missing at random" (MAR) assumption, which states that the probability for an observation to be missing can be fully explained using the values observed in the data set (for further details, see Zhu and Eisele (2013)). The PHF data are multiple imputed using the method of Rubin (1987).[1] If the MAR assumption approximately holds, and imputation models are correctly specified, the data user can be confident that the conditional distribution of the imputed variable will be well recovered by multiple imputation. The retention of general statistical features of the joint distribution of all variables is the main objective of the stochastic imputation and takes precedence over finding the most plausible value in each individual case.

Generally, a linear stochastic regression model is used to impute continuous variables (especially euro amounts). In most cases, missing values are substituted by their best linear

[1] Multiple imputation of wealth survey data was pioneered by Arthur Kennickell at the Survey of Consumer Finance (Board of Governors of the Federal Reserve System). His method is particularly well suited for surveys with complex questionnaire design and missing patterns. He allowed the PHF team to use his routines, and Cristina Barceló (EFF, Banco de España) provided a well-documented version geared to an HFCS-style survey. The PHF team is extremely grateful to them both.

predicted values, plus a normally distributed random variable. If the respondent did not report the exact value but specified an upper or a lower bound for the value, imputation is repeated until the substitute value falls into the interval. Binary variables are often indicator variables, such as the question of whether the household owns any property. They are imputed using a linear probability model. Hot deck imputation is used to impute categorical variables. Here, a missing value is replaced by an observed value of another household, which resembles the household to be imputed as much as possible in terms of the selected characteristics.

The creation of only one single imputed data set does not take into account the uncertainty of the selected imputation model and hence underestimates variances and covariances in the imputed data set. This is why the data are “multiply imputed”, by generating five different imputed data sets, or implicates. The inclusion of five data sets is a generally accepted norm. It is theoretically justified in cases where the rate of missing observations is low.

The variable **impid** indicates the number of the implicate.

5.3 Conversion

5.3.1 Time Period = Yearly

All values of variables with a flow concept (i.e. defined at specific accounting period, e.g. income) are expressed as **yearly values**. Before imputation, such variables given by the interviewed households are converted to yearly values according to reported accounting period (e.g. monthly or quarterly). All corresponding variables in the questionnaire asking for the time period to which values refer to (e.g. **dhc0510** or the sequence **dpg0300**, **dpg0310**, **dpg0320**) are not part of the SUF because they become obsolete after the conversion to yearly figures.

However, there is an important exception. The values of all variables with a flow concept in the derived variables file (**ddl2000**, **ddl2100**, **ddl2110**, **ddl2200**) correspond to **monthly** payments (or the monthly equivalent of other time frequency payments).

5.3.2 Income = Gross

The interviewed household members had the possibility to provide gross or net figures for several euro value questions. Before imputation, all net values are converted into gross values. To do this, a sophisticated algorithm is implemented to estimate the income tax and social contributions of the household members.

5.3.3 Interest Rates = Effective

Questions related to nominal interest rates (only asked if the household had no knowledge about the effective interest rate), e.g. **dhb5611**, are not provided in the SUF. Instead, the effective interest rates have been computed using the given information.

5.4 Anonymization

The Bundesbank’s raw PHF data are kept in a *formally* anonymised form, i.e. without names or addresses of households. In order to convert the PHF data into the format of a scientific

use file, various anonymization steps are performed with the objective to produce a *de-facto* anonymised data set for academic research.

In consideration of the relevant legal requirements, a number of variables are either omitted or coarsened as part of this process. The applied data protection measures are based on the current practices of factual anonymization of survey data whilst ensuring that the data's potential for analysis remains as high as possible. In addition, the panel dimension of the PHF study requires further anonymization steps, as panel studies may be subject to additional re-identification risks due to the changing characteristics of participating households, and especially household compositions, over time. This potential risk has already been proactively addressed in the first wave of the survey by making use of anonymization measures that would not have been necessary for a cross-sectional study without a panel component.

All para-data collected by the interviewers are excluded from the dataset, with the exception of the dwelling type **(dsc0100)**. Furthermore, all information and responses in text form were removed from the dataset (e.g. interviewer's comments).

Continuous variables are rounded to two digits via random rounding to offset the negative statistical effects of asymmetric rounding (To provide an example: A value of 134,400 € can be rounded either to 130,000 € or to 140,000 €). This anonymization step is thus consistent with the natural response behaviour of many households surveyed.

With regard to the regional information contained in the SUF, the anonymization steps are more extensive than the usual standards for SUFs. The risk of reducing the data's potential for analysis has intentionally been assumed for the sake of protecting the respondents' information. Nonetheless, this approach largely maintains the potential for data analysis with regard to assets and finance – the main focus of the study.

More detailed regional household information e.g. at the district level can only be obtained via on-site access at the Research Data and Service Centre or the Research Centre of the Deutsche Bundesbank in Frankfurt am Main. For more information see: https://www.bundesbank.de/en/bundesbank/research/rdsc

Additional anonymization steps include the following:

- The regional information contained in the PHF data is either deleted or coarsened (see section 6.5). Likewise, data related to demographics are coarsened. The country of birth and nationality are grouped into the categories "Germany", "euro-area countries excluding Germany", "EU excluding euro-area countries", "the former USSR excluding the Baltics", "rest of Europe including Turkey", "the Americas" and "Africa, Asia, Oceania and rest of world ". Once a person has been assigned to one of these categories, he or she remains in this country group, even after a country's inclusion in a group has changed over time (for example, when a country becomes a member of the euro area). This is necessary in order to prevent a country from potentially being re-identified. Any third nationalities are removed.

- The data concerning the relationships between the household members were consolidated into the following seven categories: “spouses”, “partners”, “parents (parents-in-law, adoptive parents, step-parents)”, “children (adoptive children, step children)”, “siblings (adoptive siblings, step-siblings)”[2]; “other relatives” and “not related”.
- The number of employees in owned businesses is consolidated into four categories (1, 2-3, 4-9, 10+). The legal forms of owned businesses are classified into four categories (“sole proprietorships”, “partnerships”, “corporations” and “other”) and the year of establishment grouped into decades.
- A stochastic error term is added to the ages of household members over the age of 70, which means that the age information in the SUF deviates within a range of -2 and +2 years from the actual age. This only marginally alters the distribution of the age data for household members over the age of 70 in the sample. In addition, the age data is top-coded at 90 years. All variables with a direct logical connection to the age of a person (e.g. year of birth) are also modified in order to maintain data consistency in accordance with the anonymization steps for age.

Furthermore, the following variables were top-coded or recoded:

- **hd0210** (“Number of self-employed private companies or businesses”): top-coded at 5
- **dhb1200f** (“Number of other vehicles - boats / yachts”): top-coded at 5
- **dhb1200g** (“Number of other vehicles – bicycles”): top-coded at 5
- **dpe9050** (“Year of birth”): wave 1 - set to 1920 (1921) if year of birth is 1920 (1921) or earlier for interview year 2010 (2011), wave 2 - set to 1924 if year of birth is 1924 or earlier, wave 3 - set to 1927 if year of birth is 1927 or earlier, wave 4 - set to 1930 if year of birth is 1930 or earlier
- **dpe1275** (“Number of children”): top-coded at 5

6 Variables, labels and codes

6.1 Identifiers

6.1.1 hhid and caseid

There are two identification numbers at the household level: **caseid** and **hhid**. **Caseid** corresponds to the **hhid** of the household where a particular person was first interviewed. They are identical for all households only in the first wave. In following waves, **caseid** and **hhid** will only be the same for newly sampled households and for panel households (see section 6.3: “Type of household”). If households split or change their composition, their **hhid** may change. However, their **caseid** stays the same over waves. Hence, **hhid** is a unique identifier.

[2] This category is not available in the first release of SUF Wave 1.

6.1.2 persid and pid

Pid is an identification number at the personal level, enumerating members within a household. Every household has a person with **pid** = 1. This person is the FKP who also answered the questions of the household level sections in almost all cases.[3] The FKP might change over waves and, hence, the **pid** of the household members is not consistent across waves. Further, it should be noted that in some cases the **pid** is not consecutive because some household members dropped out of the survey for various reasons.

A time-consistent personal identifier (**persid**) is provided that enables the tracking of specific persons over waves. **Persid** is not yet available in the first release of SUF Wave 1. However, this variable can be computed with the first release of SUF Wave 1 data in the following way: **hhid * 100 + pid**.

6.2 Missing and special codes

Missing values are indicated by one-digit negative numbers. “-3” means “question filtered”, “-1” is assigned to the interview answer “do not know”; “-2” is assigned to the interview answer “not specified/no answer”. Note that there are special codes “-4”, “-5”, “-6”, “-7” for a few questions. These can either indicate missing values, zero values, or specific answers and hence should be treated accordingly.

6.3 Type of household

In the second and third wave, three different types of households emerge: panel households (**hhart** = 1) who were already sampled in either of the previous two waves, refresher households (**hhart** = 3) stemming from the refresher sample and split households (**hhart** = 2) who have split from households originally sampled in one of the previous waves. In general, the household in which the FKP remains keeps the original **hhid** from the previous wave. Hence, for split households **hhid** and **caseid** differ. Further, a household gets a new **hhid** and is marked as a split household if its composition changes as described in the following cases:

- The FKP died and the household splits up.
- Between waves, the FKP was the only person remaining in a household with initially at least three household members, whereas all other household members changed.
- Between waves, a single person household became a large household with at least four household members and vice versa.

6.4 Sample design information

A stratified multistage sampling design is employed to select a random nationwide sample of private refresher households. The sampling design entails sampling of wealthy households with larger probability (oversampling) in order to improve the accuracy of the measurement of the wealth variables (see Eisele and Schmidt, 2013). Detailed information about the sampling

[3] In very few cases, the FKP is not a household member but an external person with the best overview of the household's finances.

of households for wave 1 is provided in Knerr et al. (2013), for wave 2 in Knerr et al. (2015), for wave 3 in Knerr et al. (2018), and for wave 4 in Knerr et al. (2022).

The sampling of new addresses for the PHF was conducted in three stages:

1) Selection of municipalities/sample points (primary sampling units = PSU)
2) Selection of street segments in large cities
3) Selection of addresses.

The sample design is reflected in the variables **stratum**, **stich** and **p_nr**.

The first stage of the sampling design divides municipalities into three strata according to their size and, for small municipalities, the share of wealthy households. The variable **stratum** contains the three first-level strata: wealthy small municipalities, other small municipalities, large cities (more than 100,000 inhabitants). Income tax statistics are used to identify small municipalities (less than 100,000 inhabitants) with a high share of wealthy households. These municipalities are oversampled at stage one.

The second stage is based on the stratification of streets in large cities. In large cities with a population of 100,000 and more, streets are grouped into two categories − streets in wealthy neighbourhoods and other streets. This information is contained in variable **stich** together with information on the stratum for small municipalities. Wealthy street sections are identified based on micro-geographical information (e.g. quality of residential area, type of dwelling, purchasing power indicator) and are oversampled.

In the third stage, addresses of adults persons older than 18 are drawn from a public register out of the resulting four final strata. The sample points are equally large address clusters. Sample point numbers are provided in the variable **p_nr**. Households having the same **p_nr** are located nearby. Besides this, the value of **p_nr** does not contain any regional information.

6.5 Geographical information

As mentioned in section 5.4, regional information is coarsened for anonymisation purposes. The Federal states are grouped into the following four categories (variable **bland**): “north” (Bremen, Hamburg, Lower Saxony and Schleswig-Holstein), “south” (Baden-Württemberg, Bavaria and Hesse), “west” (North Rhine-Westphalia, Rhineland-Palatinate and Saarland) and “east” (Berlin, Brandenburg, Mecklenburg-West Pomerania, Saxony, Saxony-Anhalt and Thuringia).

The categories on municipality size (variable **polgk**) are reduced from seven to five (categories 1 and 2 as well as 4 and 5 are combined) with gradations at 5,000, 20,000, 100,000 and 500,000 inhabitants.

The original ten BIK categories[4] are reduced to five categories. Categories 5 and 7, 6 and 8, as well as 1 to 4 are combined. The five categories of the BIK regional size class (variable

[4] They were first introduced by the research institute “BIK Aschpurwis + Behrens GmbH”. BIK category 1: < 2,000 inhabitants, BIK category 2: 2,000 – 4,999 inhabitants, BIK category 3: 5,000 – 19,999 inhabitants, BIK category 4: 20,000 – 49,999 inhabitants, BIK category 5: 50,000 – 99,999 inhabitants and BIK structure type 2 (city region)/ 3 (second tier towns)/ 4(third

bikgk) are: < 50,000 inhabitants; 50,000-499,999 inhabitants and BIK structure type 2/3/4; 50,000-499,999 inhabitants and BIK structure type 1; ≥ 500,000 inhabitants and BIK structure type (2/3/4); ≥ 500,000 inhabitants and BIK structure type 1.

For wealthy small municipalities in region “east”, values for **polgk** as well as **bikgk** are missing due to anonymisation purposes.

7 Flag Variables

The variable name of a flag variable is composed of the name of the related variable and the suffix “fl”.

7.1 Form

The flag codes have four digits with the exception of the two standard cases "not applicable" (flag 0) and "recorded as collected" (flag 1) as well as a two digit flag (flag 12). The two digit flag indicates that values were deliberately not collected. This is for example the case, if data originally stems from the previous wave (preload data).

7.2 Structure of 4-digit flags

The first digit of the flag code indicates whether the value had been imputed or not:
1=not imputed, 2=imputed.

The second digit reveals whether the value had been edited:
0=not edited, 1=edited, 2=set to missing or deleted

The third digit informs about the editing status and reason for editing:
0=not edited, 1=implausible value, 2=recoded because answer contains text, 3=currency conversion, 4=time-period conversion, 5=net-gross conversion, 6=other recoding, 7=set to missing due to editing a preceding (head) variable, 9=reason for editing in wave 1 not identifiable (applies only to flag variables in wave 1)

The fourth digit indicates the original status of the value:
0=original value “don’t know”, 1=original value “no answer”, 2=missing value due to a missing answer in a preceding question, 3=value collected, but edited/deleted/recorded (reason stated in the third digit), 4=value had been provided in interval ranges, 5=CAPI-error or interviewer-error, 6=value was consciously not collected.

Thus, imputed values are those with a flag variable value greater equal 2000. Consider that many missing values (value -3 / question filtered) also have a corresponding imputation flag of 2002 due to missing values in preceding questions.

7.3 4-digit flags in the first release of SUF Wave 1

The first release of SUF Wave 1 differs in the meaning of the four digit flags with regard to the second, third, and fourth digit. However, this is not the case for consecutive versions of

tier towns), BIK category 6: 50,000 to 99,999 inhabitants and BIK structure type 1 (metropolitan area), BIK category 7: 100,000 – 499,999 inhabitants and BIK structure type(2/3/ 4), BIK category 8: 100,000 – 499,999 inhabitants and BIK structure type 1, BIK category 9: ≥ 500,000 inhabitants and BIK structure type (2/3/4), BIK category 10: ≥ 500,000 inhabitants and BIK structure type 1.

the SUF for the first wave, as their flags have been adjusted to be consistent with the meaning in the other waves. The structure of four digit flags in the first release of SUF Wave 1 is as follows:

The first digit of the flag code indicates whether the value had been imputed or not:
1=not imputed, 2=imputed.

The second digit indicates whether and how the value had been edited:
0=not edited, 1=manually edited, 2=manually set to missing, 5=automatically edited, 6=automatically set to missing

The third digit is used for purely internal purposes.

The fourth digit indicates the status of the value:
0=original value “don't know”, 1=original value “no answer”, 2=missing value due to a missing answer in a preceding question, 3=implausible value, 4=value had been provided in interval ranges, 5=CAPI-error or interviewer-error, 6=recoded value, 7=currency conversion, 8=net-gross-conversion.

8 Weights

8.1 Cross-sectional household weights

The weights are constructed over multiple stages. First, design weights are assigned to correct for unequal selection probabilities and the oversampling of wealthy households that result from the complex sample design. Second, these weights are adjusted for non-response using estimated response propensities from a logit response propensity model. The weights of households with lower response propensity are inflated accordingly. In order to contain the variance of the weights, these factors are trimmed.

In the final step, the weights are calibrated to ensure that weighted estimates accurately represent the population in important dimensions not captured by sample design. In order to match the overall marginal distributions, calibration relies on external information provided by the German Microcensus 2010, 2014, 2016, and 2019. Some of the marginal distributions used in calibration are either referring to the household structure, others to the status of the household’s main income earner. The former group comprises household size, region (“Federal States”), municipality size (“politische Ortsgrößenklasse”), ownership status of main residence, and size of main residence for owners. The group of variables referring to the main income earner consists of labour market status, nationality and combinations of age with gender and the highest schooling degree, respectively.

Final (calibrated and adjusted) household weights are provided in the variable **exhoch_hh**. This expansion factor indicates how many households in Germany are represented by the individual household. The household weight **exw_hh** builds on **exhoch_hh** but is adjusted to have the mean value 1.

8.2 Longitudinal household weights

The longitudinal weights **wlong** for wave 2 are constructed at household level and are only available for panel households (i.e. the split households are not included).

They were constructed as follows:

- The final weights of wave 1 serve as starting base weights and are first adjusted for unit non-response of panel households in wave 2 (using the already estimated response propensities of panel households from the construction of the cross-sectional weights).
- Second, the resulting weights for panel households are calibrated to wave 1 statistics (using the wave 1 data of all responding households).

As a result, the characteristics of the panel households at wave 1 using the longitudinal weights are similar to the characteristics of the whole sample at wave 1, using the cross-sectional weights. Longitudinal-household weights for the third wave are not yet available.

Note that there are various possible methodologies for constructing longitudinal weights. We will continue to research in that field and may revise our methodology.

8.3 Replicate Weights

The w-file provides bootstrap replicate weights in order to enable efficient **variance estimation** even if, because of disclosure control reasons, not all important design elements can be passed on to the research community. The replicate weights are the result of bootstrap simulations that take the sample design features into account. Similarly to the ordinary household weights **exhoch_hh**, the weights are adjusted for non-response and calibrated.

9 Interim Surveys

Interim surveys serve to stay in contact with households and to gather information on current topics of interest. As an example, the 2019 PHF interim survey included a number of questions on fintech. Intermediate surveys are much shorter than the main surveys and were conducted by infas by post.

The interim survey conducted in 2020 served mainly to bridge the period between the last main survey on household wealth in Germany in 2017 and the main survey postponed to 2021 as a result of the pandemic. The Bundesbank's Research Centre, together with infas, conducted a postal interim survey, which contained specific questions on the impact of the coronavirus pandemic on households and their saving behavior. Information on the asset structure was also collected, but not as detailed as in the main surveys. The different survey mode of the interim survey (post versus face-to-face in the main survey) and the resulting difference in the design of the questions on wealth further restrict the comparability of the interim survey with the main waves. In particular, it is not possible to compare the magnitude of absolute assets consistently over time. However, the survey results provide an insight into the impact of the pandemic on households' finances and wealth distribution. A total of 4550

households participated in the interim survey, with the majority already participating in previous surveys.

References:

ECB (2015) Using the HFCS with STATA. Household Finance and Consumption Survey Technical Series. Version 1.5.

Knerr, P, Chudziak, N, Kleudgen, M and Steinwede, A (2022) Methodenbericht Private Haushalte und ihre Finanzen (PHF), 4. Erhebungswelle 2021, anonymisierte Fassung https://www.bundesbank.de/resource/blob/825518/acefbf448b4b1c9f6c61c26813890d0a/mL/methodenbericht-welle4-data.pdf

Knerr, P, Aust, F, Chudziak, N, Gilberg, R and Kleudgen, M (2018) Methodenbericht Private Haushalte und ihre Finanzen (PHF), 3. Erhebungswelle 2017, anonymisierte Fassung https://www.bundesbank.de/resource/blob/798122/17644354b20fb4d450f7267e91444fbb/mL/methodenbericht-welle3-data.pdf

Knerr, P, Aust, F, Chudziak, N, Gilberg, R and Kleudgen, M (2015) Methodenbericht Private Haushalte und ihre Finanzen (PHF), 2. Erhebungswelle 2014, anonymisierte Fassung https://www.bundesbank.de/resource/blob/617398/690b0a49a963e4b71ca572fb1d838603/mL/methodenbericht-welle2-data.pdf

Knerr, P, Chudziak, N, Gilberg, R and Kleudgen, M (2013) Methodenbericht Vermögenssurvey, 1. Erhebungswelle 2010/2011, anonymisierte Fassung https://www.bundesbank.de/resource/blob/617402/40c963b68419952a17548a31179276a3/mL/methodenbericht-welle1-data.pdf

Rubin, D B (1987). Multiple Imputation for Nonresponse in Surveys. New York, Wiley.

Schmidt, T, and Eisele, M (2013) Oversampling vermögender Haushalte im Rahmen der Studie "Private Haushalte und ihre Finanzen (PHF)“. https://www.bundesbank.de/resource/blob/617146/92b765b0227577b65ccec0fdbd8a86bc/mL/phf-oversampling-data.pdf

Von Kalckreuth, U, Eisele, M, Le Blanc, J, Schmidt, T and Zhu, J (2012) The PHF: a comprehensive panel survey on household finances and wealth in Germany, Bundesbank Discussion paper 13/2012. https://www.bundesbank.de/resource/blob/617148/e6ba33aa5cb2f0fd12db5a4205b5cd23/mL/2012-07-10-dkp-13-data.pdf

Zhu, J and Eisele, M (2013) Multiple imputation in a complex household survey - the German Panel on Household Finances (PHF): challenges and solutions. https://www.bundesbank.de/resource/blob/617166/721c350ca585dc45f3613a86a2bbbb32/mL/phf-imputation-data.pdf