SOEP The German Socio-Economic Panel study

341

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP — The German Socio-Economic Panel study at DIW Berlin 2016

SOEP FiD – 'Familien in Deutschland', Data Documentation Release FiDv4.0

Mathis Fräßdorf, Rainer Siegers, Stefan Damerow, Moritz Mannschreck, Guido Putzke, Alexander Raith, Nina Scherner, Juliana Werneburg, Linda Wittbrodt, Malisa Zobel

DIW SOEP

Running since 1984, the German Socio-Economic Panel study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing. The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at **http://www.diw.de/soepsurveypapers**

Please cite this paper as follows:

Mathis Fräßdorf, Rainer Siegers, Stefan Damerow, Moritz Mannschreck, Guido Putzke, Alexander Raith, Nina Scherner, Juliana Werneburg, Linda Wittbrodt, Malisa Zobel. 2016. SOEP FiD – 'Familien in Deutschland', Data Documentation Release FiDv4.0. SOEP Survey Papers 341: Series D. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)
DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany
soeppapers@diw.de

SOEP FiD – 'Familien in Deutschland', Data Documentation Release FiDv4.0

Mathis Fräßdorf, Rainer Siegers, Stefan Damerow, Moritz Mannschreck, Guido Putzke, Alexander Raith, Nina Scherner, Juliana Werneburg, Linda Wittbrodt, Malisa Zobel

February 2014

The dataset FiDv4.0 is released with doi: 10.5684/soep.fid.v4.0. The data of the FiD-Study is now integrated completely in SOEP-Core, Version 31ff. (10.5684/soep.v31). The households will continue to be monitored as samples L1-L3 and interviewed with the SOEP-Core questionnaire.

Preface

This documentation consists of various parts and is meant to guide the user through the latest release of the "Familien in Deutschland" FiD data collection. This short introduction provides an overview of files and concepts, while the specific documentation files give some insights on the data generation process.

In case of questions or comments, please feel free to contact any one from the FiD-Team at the SOEP division of the DIW Berlin. We thank all SOEP researchers for their support and guidance throughout FiD's first three waves of data collection, especially Elisabeth Liebau for help during the questionnaire construction phase, Martin Kroh for help on the weighting procedures, and Juliana Werneburg for her help on the marital datasets. Special thanks are in order for our current and past student assistants Stefan Damerow, Moritz Mannschreck, Guido Putzke, Alexander Raith, Nina Scherner, Linda Wittbrodt and Malisa Zobel for their excellent assistance throughout the data generation and documentation. We are also grateful to the SOEP-FiD group at TNS-Infratest, who provided valuable input to the documents. This work would not have been possible without the financial support of the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and the Federal Ministry of Finance (BMF). Last, but by no means least, we especially thank the FiD respondents for entrusting their information to us.

Berlin, February 2014

Mathis Fräßdorf

Jürgen Schupp	jschupp@diw.de	030/89789-238
Rainer Siegers	rsiegers@diw.de	030/89789-239
C. Katharina Spieß	kspiess@diw.de	030/89789-254

Contents

General Documentation

Basics, Structure, and Overview of Datasets

Mathis Fräßdorf (geb. Schröder)

Contents

General information

The project "Familien in Deutschland" – "Families in Germany" – is a longitudinal panel study financed by the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and the German Federal Ministry of Finance (BMF). Its main purpose is to provide researchers with new and better data on specific groups in the German population: low-income families, families with more than two children, single parent families as well as families with young children. The data are the backbone of the first large scale evaluation of family policy measures in Germany on behalf of the two involved ministries.

The purpose of this documentation is to provide the user with the essentials of the data from "Familien in Deutschland" (FiD). We briefly describe the samples, the data structure, the naming logic, and give some basics on combining different data sets. We then turn to the different files included in the current data distribution (version 4.0), and for each provide a quick overview of the most important facts. More information is available from an Article to be published in "*Schmollers Jahrbuch*" (2013, 133, 4) by Schröder, Siegers & Spieß or in the SOEP-Survey paper 556[1]. Either publication should be cited when using the FiD data.

This documentation is the start to the analysis with the FiD data – it somewhat resembles the *Desk Top Companion* known from the SOEP data collection, although it has been specifically constructed to suit the FiD data. There are several file-specific documentation files which are included as separate files in the data distribution. Users should refer to them, as they dive into the details of each data set. The questionnaire files, which are available as separate pdf-files for each questionnaire in the field, complement the different documentations. As FiD has a large overlap with the SOEP in terms of questions and variables, a further very useful documentation tool is provided in the correspondence file (FiD-SOEP_Correspondence_v4.0.xlsx), where variable correspondences are given within FiD (i.e. across waves) and across FiD and SOEP (i.e. variable correspondence from 2010 – "ba", 2011 – "bb", and 2012 – "bc"). This allows to 'translate' from one wave or dataset to the other. In addition, a future online documentation for the SOEP (formerly known as SOEPinfo) will include FiD and other related studies as well.

[1] see Schröder, Siegers & Spieß, 2013, included in the data distribution and also available online at: http://www.diw.de/documents/publikationen/73/diw_01.c.421623.de/diw_sp0556.pdf)

The very basics of "Familien in Deutschland"

The role model for this project is the well-known "Socio-economic Panel" (SOEP), a representative study of the German population which runs annually since 1984. FiD is shaped after the SOEP in terms of structure (i.e. a household sample with individual interviews) and questionnaire content. However, the sample was drawn and weighted to represent the following four groups of the German population:

1. Families in "critical income brackets"
2. Single parents
3. Families with more than two children
4. Cohorts of children born between 2007 and 2010

The first three samples were selected through a screening process, which defined eligibility for households through a quick telephone interview before the actual interview. These three samples are thus often referred to as the "Screening Sample 2010". The fourth group was selected through random sampling from German registries, and is referred to as the "Cohort Sample".

All interviews were conducted in face-to-face mode, i.e. an interviewer was present during the interview. The general mode of the interviews was CAPI (Computer Assisted Personal Interview), however, some questionnaires allowed for a PAPI (Pen and Paper Interview) mode as well.

In 2010, wave 1 started containing overall 4.574 interviewed households with a total of 17.002 individuals. There are a total of 17.352 different interviews available, which include household interviews, person interviews and proxy interviews for children living in the household. Especially the latter provide valuable information for the evaluation purposes. The longitudinal dimension started with wave 2 in 2011. Here an additional subsample 924 households of both single parents and families with more than two children was added to the study, called the "Screening Sample 2011".[2]

[2] For more details on the sampling procedures and definitions, on response rates, and on the number of observations, please see the methodological reports by TNS Infratest at: http://www.diw.de/de/diw_01.c.405871.de/fid_dokumentation.html.

Data structure: types of data files, names and data organization

"Familien in Deutschland" (FiD) is an annually conducted panel study, providing the collected information in various different data files. Overall, there are 100 files included in the data distribution of February 2014 (version 4.0), which can come on one of three different levels: household, adults, and children. It is always possible to match data from one level with the other levels via identifiers common to all data files (see the next section for details). The different levels are one reason for storing information in separate files rather than in one large data file. However, we do not combine data within one information level, either, because such a combination would make it very cumbersome to deal with the data – not only from a computational point of view, but keeping an overview of content and datasets with several hundreds of variables might not be easy.

In general, FiD publishes two different types of datasets: original files (i.e. datasets following the structure of questions and not including any generated variables) and generated files (datasets produced by using information from one or more of the original files).[3] Original files are cross-sectional, while generated files may be cross-sectional or longitudinal. It is easy to identify cross-sectional datasets in FiD: they all have a wave specific label, which in documentation files, we generally denote by the letter "$". The "$" is a substitute for a prefix, e.g. "f10" corresponds to the data collection in 2010. Cross-sectional datasets usually have only one observation for each key unit (person or household), while longitudinal datasets may have multiple observations at different time points for each key unit. Longitudinal generated data files can be either annually based (i.e. each year is an observation) or spell based. We provide more detail on the structure of each dataset below (see "FiD Data Files").

Wave identifiers ("$") are generated in a simple way: Each wave is abbreviated by using the lower-case letter "f" and the third and fourth digit of the respective year in which the survey took place. For example, the first wave is called "f10", the second wave is called "f11", and so on. Wave identifiers are always used as a prefix to a dataset, and should not be confused with the "$$" used as suffix for variables in some datasets (mainly ***$pgen*** and ***$hgen***). In these files, the double dollar sign "$$" indicates a wave specific variable – it is a substitute for the last two digits of the respective year (e.g. SIZE$$ would be found as SIZE10 in ***f10hgen***).[4]

[3] There is one exception to this rule with the dataset ***$kind***, see below.

[4] As a matter of convenience to increase readability, throughout the documentation we refer to variable names in CAPITAL letters, even though in the datasets, data are usually in lower case letters. Datasets are in general referred to in ***bold italic*** letters.

For each generated data file, FiD provides an additional detailed documentation on the specifics of the generating process. With regard to the original data files, no additional documentation exists, as variables in the original files are named according to the sequence of the questionnaire. Principally, the following rule has been used for naming variables in original data files: the wave specific label is followed by the first letter of the name of the data file, followed by three digits indicating the number of a specific question in the questionnaire. Imagine, for example, that you are interested in question 10 of the "Household questionnaire" (i.e. "equipment of apartment") and you want to find the corresponding variable with information on that question in wave 1 (2010).

This is how you would proceed:

- Current wave: "f10" +
- Household Questionnaire: "h" +
- 3-Digit Nr.: "010"

Table 1: Overview of Original Data Files and Corresponding Questionnaires

Information on	**Dataset**	**Questionnaire Name (pdf-file)**
Person Level (Adult)	$p	$Person_Qn (except Questions with prefix "L")
	$lela	$Person_Qn (all Questions with prefix "L")
	$pkal	$Person_Qn
	$jugend	$Youth_Qn
Person Level (Child)	$eltern1	$Parent_Qn1
	$eltern2	$Parent_Qn2
	$eltern3	$Parent_Qn3
	$eltern4	$Parent_Qn4
	$eltern5	$Parent_Qn5
	$eltern6	$Parent_Qn6
Household Level	$h	$Household_Qn

The variable you are searching for is variable F10H010. Looking for it in the dataset ***f10h***, you find that there are 13 variables called F10H010, each having a suffix. Going back to the questionnaire you can see that there are 13 answer possibilities (so-called items) for the "equipment of apartment"-question. Variables that correspond to questions of this type (i.e.

more than one answer is possible) generally receive an "A" for the first item, a "B" for the second and so forth. Hence, by taking a look into the respective questionnaire it is easily possible to match each variable with the corresponding question and vice versa. There is a slight exception to the above construction rule: in case of the "Parent-Questionnaires", the 3-digit Number refers to the questionnaire number (i.e. "1" to "6") in the first digit, while the last two identify the question number. Thus, a variable called f10e104 would indicate to look for Parent-Questionnaire 1 and the fourth question (i.e. you should read: f10-e1-04).[5]

Combining data files

Sometimes it is necessary to combine information from various datasets. For example if someone working with the generated ***bioage*** data files (children specific information) wants to have more information on mothers, she needs to take that information from another file (e.g. the p-files). This can be easily done by merging the respective datasets. The variables of the using dataset are then added to the already existing observations in the master dataset. When merging data files it is crucial to know which variable uniquely identifies observations in the master and using file. Such a key variable is called an identifier and varies from dataset to dataset.

Usually the best way to merge is to use datasets which use the same logical observation to organize the data. For example, if you were interested in household income (e.g. ***$hgen*** dataset), and wanted to add information on whether a household has an immigrant background or not, you would need to merge the ***$hgen*** dataset with the ***$hbrutto*** dataset. The identifier to match the two would be the original household identification number (HHNR) or the current household identification number (HHNRAKT). But even if the logical observation organizing the data is a person and you are interested in some information based on household, you are able to match each person with the corresponding household by specifying the household identifier as the key variable.

It should be stressed that if your information is based on the household level and you want to keep it that way, you should refrain from merging it with information based on the person level (since a household includes various respondents and observations thus cannot be uniquely identified). Nevertheless, if you want to keep the household as your logical

[5] Some known exceptions should be noted: In the biography section of the person questionnaire, the questions L62 to L62d have been coded as $L062A to$L062E. In the household questionnaire for 2010, question 42b has been coded as F10H043, and question 43 has been named F10H044. For the household questionnaires starting with wave 2 (2011) exceptions are questions W1, W2, W3, W4, and W5, which are named $H001A, $H001B, $H001C, $H001D, and $H001E. Additionally, question 44b is coded as $H044Z.

observation, but still want use information from a dataset based on person level, you can do so by first collapsing the person-based information to the household level.

Table 2 Overview of identifiers

Household Identifiers	**Person Identifiers**
HHNRAKT	PERSNR
$HHNR	PERSNRM, PERSNRRESP
HHNR	COUPID
	PARTNO$$
	FATHNO$$/MOTHNO$$

As a further complication, datasets on the person level can either have adults or children as the logical observation organizing the data, with the identifiers PERSNR and PERSNRM or PERSNRRESP. It is important to keep this in mind when merging datasets. Going back to the example above, consider augmenting the ***bioage01*** data files with more specific information on the mothers of these children. The ***bioage01*** data files are organized based on children, thus PERSNR refers to a specific child. Parents, or in our example, mothers can be identified using the variable PERSNRRESP. Hence a merge of ***bioage01*** to ***f10p*** using PERSNR would lead to no matching observations, as children are not included in ***f10p***. To achieve the merge, PERSNR in either ***f10p*** or ***bioage01*** would have to be renamed (e.g. rename PERSNR in ***bioage01*** to PERSNRK, then rename PERSNRRESP to PERSNR, and merge ***f10p*** via PERSNR.) Similarly, children can be matched to their parents by using either FATHNO$ (PERSNR of the father in ***$kind***) or MOTHNO$ (PERSNR of the mother in ***$kind***) and partners can be matched by using PARTNO$ (PERSNR of the partner in ***$pgen***).

FiD data files

Basic Data Files

Basic data files are those which allow the user to retrieve basic information on the unit of observation. ***ppfad*** and ***hpfad*** are datasets with which the user can monitor the development of each person and household through the panel life, along with some generated information. The ***$pbrutto*** and ***$hbrutto*** files provide similarly important information about the interviewing process for each wave. ***hbrutt10_fid*** and ***hbrutt11_fid*** contain the gross sample with which the survey started, i.e. the Cohort and Screening Gross Sample in 2010 and the Screening Gross Sample in 2011.

ppfad

Sample	all individuals in FiD
Unit of Observation	Person
Identifier	PERSNR (HHNR, $HHNR)
Description	Generated person level path variables (basic demographic variables) follow a person's path throughout her panel life.
Documentation	Documentation ***ppfad***

hpfad

Sample	all participating households in FiD
Unit of Observation	Household
Identifier	HHNR ($HHNR)
Description	Generated household level path variables, follows a household's path throughout its panel life.
Documentation	Documentation ***hpfad***

$hbrutto

Sample	all participating households in FiD
Unit of Observation	Household
Identifier	HHNR (HHNRAKT, $HHNR)
Description	Information about all participating households in the respective wave
Documentation	Documentation ***hbrutto***

$pbrutto

Sample	all individuals in FiD
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	Information about all individuals in the respective wave
Documentation	Documentation ***pbrutto***

hbrutt10_fid

Sample	all households selected to participate in FiD in 2010
Unit of Observation	Household
Identifier	HHNR (HHNRAKT, $HHNR)
Description	Information about all households selected to participate in FiD in 2010 for the first time, i.e. the screening population of low income families, single parents, and families with more than two children as well as the cohort sample. This gross sample file contains information for non-responding households as much as possible.
Documentation	There is no specific documentation, but the documentation on ***hbrutto*** will help the user.

hbrutt11_fid

Sample	all households selected to participate in FiD in 2011
Unit of Observation	Household
Identifier	HHNR (HHNRAKT, $HHNR)
Description	Information about all households selected to participate in FiD in 2011 for the first time, i.e. the screening population of single parents and families with more than two children. This gross sample file contains information for non-responding households as much as possible.
Documentation	There is no specific documentation, but the documentation on ***hbrutto*** will help the user.

Original Data Files

Original data files are those, which are left unchanged and do not include any generated variables (see section 1 for a more specific explanation). All generated data files are built up from the original data files and can be reproduced in that way. Nonetheless, we do not distribute the answers from open-ended questions in the original data files for data protection reasons. As a matter of fact, no string variables are included in the data, except for the data in ***$pkal***. However, string data are coded into categories if possible and are then included as numerical values in the generated files.

$h

Sample	all households participating in the survey
Unit of Observation	Household
Identifier	HHNR, HHNRAKT, $HHNR
Description	Annual questions concerning the household as a whole
Documentation	$Household_Qn

$p

Sample	adults turning 18 in the current survey year or older
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	Annual individual questions
Documentation	$Person_Qn (without L-questions)

$lela

Sample	adults turning 18 in survey year or older, answering for the first time
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	Biographical person questions that are part of the Personal Questionnaire, questions are named by L#. In FiD 2010 there was one biographical part in questions L1 to L67. In 2011, a second part was added with questions L70 to L100, with the goal that persons participating for the first time in 2011, receive part 1, revisited respondents receive part 2, so that within two waves, a complete biography is gathered, that resembles the SOEP questionnaire. However, due to a technical error, all sample members of the samples drawn in 2010 received part 2, even if they entered the study only in 2011. The variable $LELTYP specifies which part of the biography questionnaire was filled out.
Documentation	$Person_Qn (questions labeled by L#)

$pkal

Sample	adults turning 18 in survey year or older
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	Provides the calendar information from the personal questionnaire on what the respondent did in each month last year. Information is recorded in string variables.
Documentation	$Person_Qn, activity calendar

$jugend

Sample	all persons turning 17 in the survey year
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	The youth questionnaire is the first questionnaire a FiD member will ever fill out herself. Note that if a youth questionnaire has been conducted, no biography part will be conducted with this person in the future.
Documentation	$Youth_Qn

$eltern1

Sample	newborn children or turning 1 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by the mother or single parent father living in the household.
Documentation	$Parent_Qn1

$eltern2

Sample	children turning 2 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by the mother or father living in the household.
Documentation	$Parent_Qn2

$eltern3

Sample	children turning 3 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by the mother or father living in the household.
Documentation	$Parent_Qn3

$eltern4

Sample	children turning 6 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by the mother or father living in the household.
Documentation	$Parent_Qn4

$eltern5

Sample	children turning 8 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by both the mother AND the father if living in the household. Note that this leads to multiple observations per child. PERSNR and PERSNRM can be used to identify a single observation.
Documentation	$Parent_Qn5

$eltern6

Sample	children turning 10 in the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRM (HHNR, HHNRAKT, $HHNR)
Description	Proxy parent questionnaire about children living in the household. Answered by both the mother AND the father if living in the household. Note that this leads to multiple observations per child. PERSNR and PERSNRM can be used to identify a single observation.
Documentation	$Parent_Qn6

$kind

Sample	all children turning at most 16 during the survey year
Unit of Observation	Person (child)
Identifier	PERSNR, FATHNO10, MOTHNO10 (HHNR, HHNRAKT, $HHNR)
Description	This file contains mainly original data from the so-called “children’s matrix” in the household questionnaire. The household head answers questions about all children living in the household. The data – originally on the household level – are then converted into children level information. Added to this data are the identifiers for mother and father as well as the type of their relationship (i.e. biological, social, or adoptive).
Documentation	$Household_Qn, Documentation ***$kind***

$luecke

Sample	all eligible persons of the previous year who did not participate

Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT, $HHNR)
Description	If former FiD respondents decide not to participate in a given year, the following time the household is visited they are again asked to participate in the study. In such a case, they fill out the regular person questionnaire, but are also interviewed about their year of absence with a short questionnaire containing the main information. The information of this short questionnaire is stored in ***$luecke*** (from German "Lücke", gap), containing the exact same variable names of the previous year's p-questionnaire. The data from this file belong to the previous wave and are hence stored in that year's folder. (E.g. the data gathered in 2012 from respondents not participating in 2011 is called ***f11luecke***, and this file is stored in the folder ~/2011.) Also, in ***ppfad*** the $NETTO code of the previous wave is changed accordingly (i.e. from 130 to 131).
Documentation	$Luecke_Qn

Generated Data Files

Generated data files are created for the user to ease the handling of the data. For the most part, they are constructed from original data which are also included in the data distribution. However, some files use information which for data protection reasons cannot be given to the user directly, for example detailed neighborhood information, which is used in the construction of the weights.

$pgen

Sample:	all adults turning 17 in survey year or older
Unit of Observation	Person
Identifier:	PERSNR (HHNR, HHNRAKT)
Description:	Generated person-level variables
Documentation:	Documentation on ***pgen***

$hgen

Sample:	all households participating in the survey
Unit of Observation	Household
Identifier:	HHNR, HHNRAKT
Description:	Generated household-level variables
Documentation:	Documentation on ***hgen***

artkalen

Sample	all adults turning 18 in survey year or older.
Unit of Observation	Person
Identifier	PERNSNR, SPELLNR (HHNR)
Description	The file contains monthly spells for events starting in January 2009, which are generated from the activity calendar ***$pkal***. This is in contrast to ***pbiospe*** (see below), where spells are in yearly durations, and events prior to the first wave are included. ***artkalen*** is longitudinal, as it combines information from all activity calendars. Observations in the data are uniquely identified through PERSNR and SPELLNR.
Documentation	There is no specific documentation; the activity calendar in $Person_Qn provides some help

pbiospe

Sample	all adults turning 18 in survey year or older.
Unit of Observation	Person
Identifier	PERNSNR, SPELLNR (HHNR)
Description	The file contains yearly spells for events starting at the age of 15 for each respondent, generated from the biographical activity calendar

	collected in the second part of the biography ($LELTYP=2 or 3). While the biography is recorded only once, ***pbiospe*** is updated every year with spell information from ***artkalen***. ***pbiospe*** is longitudinal, as it combines information from biography and all activity calendars. Observations in the data are uniquely identified through PERSNR and SPELLNR.
Documentation	Documentation on ***pbiospe***

phrf and phrf_fidsoep

Sample	all participating persons
Unit of Observation	Person
Identifier	PERSNR (HHNR, $HHNR)
Description	Person weighting factors for each wave, also includes the inverse probabilities of staying in the sample. ***phrf_fidsoep*** provides this information for the integrated SOEP-FiD data.
Documentation	Documentation on ***phrf*** and ***hhrf***

hhrf and hhrf_fidsoep

Sample	all participating households
Unit of Observation	Household
Identifier	HHNR, $HHNR
Description	Household weighting factors for each year, also includes the inverse probabilities of staying in the sample. ***phrf_fidsoep*** provides this information for the integrated SOEP-FiD data.
Documentation	Documentation on ***phrf*** and ***hhrf***

$mipinc

Sample	all persons participating in the survey
Unit of Observation	Person
Identifier	PERSNR, _MJ (HHNR, HHNRAKT)
Description	Contains multiply imputed data on personal incomes. In this specific dataset observations are uniquely identified by the combination of PERSNR with the identifier for each imputed implicate, _MJ.
Documentation	Documentation on ***mihinc*** and ***mipinc***

$mihinc

Sample	all households participating in the survey
Unit of Observation	Household
Identifier	HHNR, _MJ, HHNRAKT
Description	Contains multiple imputed data on household incomes and other payments. In this specific dataset observations are uniquely identified by the combination of HHNR or HHNRAKT with the identifier for each imputed implicate, _MJ.

Documentation	Documentation on ***mihinc*** and ***mipinc***

bioage01

Sample	newborn children (aged 0-1 during survey year)
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)
Description	Contains information about children when they were 0-1 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children are born into this age group.
Documentation	Documentation on ***bioage*** files

bioage02

Sample	children aged 1-2 during survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)
Description	Contains information about children when they were 1-2 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children grow into this age group.
Documentation	Documentation on ***bioage*** files

bioage03

Sample	children aged 2-3 during survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)
Description	Contains information about children when they were 2-3 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children grow into this age group.
Documentation	Documentation on ***bioage*** files

bioage06

Sample	children aged 5-6 during survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)
Description	Contains information about children when they were 5-6 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children grow into this age group.
Documentation	Documentation on ***bioage*** files

bioage08p1 and bioage08p2

Sample	children aged 7-8 during survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)

Description	Contains information about children when they were 7-8 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children grow into this age group. Note that ***bioage08p1*** contains all children for which a questionnaire was filled out. ***bioage08p2*** adds information from an additional parent who has filled out the questionnaire.
Documentation	Documentation on ***bioage*** files

bioage10p1 and bioage10p2

Sample	children aged 9-10 during survey year
Unit of Observation	Person (child)
Identifier	PERSNR, PERSNRRESP (HHNRAKT)
Description	Contains information about children when they were 9-10 years old and living in the household. This file is not longitudinal, but contains multiple cross sections as children grow into this age group. Note that ***bioage10p1*** contains all children for which a questionnaire was filled out. ***bioage10p2*** adds information from an additional parent who has filled out the questionnaire.
Documentation	Documentation on ***bioage*** files

bioagel

Sample	all children in bioage01-bioage10
Unit of Observation	Person (child)
Identifier	PERSNR, BIOAGE, PERSNRRESP (HHNRAKT)
Description	Contains the combined information about children who were covered in the ***bioage*** files in a longitudinal format, i.e. multiple observations per child are possible.
Documentation	Documentation on ***bioage*** files

Biobirth

Sample	all adults at least 17 years old
Unit of Observation	Person
Identifier	PERSNR (HHNR)
Description	Contains information about a respondent's biological children and the situation surrounding the birth. This information is mainly based on the biography part of the person questionnaire.
Documentation	Documentation on ***biobirth***

biomarsy

Sample	all adults at least 17 years old
Unit of Observation	Person
Identifier	PERSNR SPELLNR (HHNR)
Description	Contains yearly spell information on the respondent's martial history. This information is mainly based on the biography part of the person

	questionnaire. Unique observations are identified through PERSNR and SPELLNR.
Documentation	Documentation on ***biomarsy*** and ***biocouply***

biocouply

Sample	all adults at least 17 years old
Unit of Observation	Person
Identifier	PERSNR SPELLNR (HHNR)
Description	Contains yearly spell information on the respondent's partner history, including information on marriages. This information is mainly based on the biography part of the person questionnaire. Unique observations are identified through PERSNR and SPELLNR.
Documentation	Documentation on ***biomarsy*** and ***biocouply***

bioresid

Sample	all adults at least 17 years old
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT)
Description	Contains information on when a person moved into her current accommodation, and whether secondary accommodations exist. The information so far stems from the biography questionnaire and information on moves during the panel.
Documentation	No documentation yet, please refer to the SOEP documentation for this dataset.

bioparen

Sample	all adults at least 17 years old
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT)
Description	Contains information on a person's parents, mainly information from the second part of the biography questionnaires. This dataset contains the same variable names as the corresponding SOEP data, and can thus be used easily in combination with these data.
Documentation	Documentation on ***bioparen***

bioage17

Sample	youths turning 17 years during the survey year
Unit of Observation	Person
Identifier	PERSNR (HHNR, HHNRAKT)
Description	Contains information from the youth questionnaires (***$jugend***) in a combined format, i.e. all persons who ever answered this questionnaire are included in this dataset (similar to the ***bioage01*** - ***bioage06*** files).
Documentation	Documentation on ***bioage17.***

paradatal

Sample	all respondents
Unit of Observation	Questionnaire
Identifier	SYEAR PERSNR PERSNRK, $QSTNR (HHNR, HHNRAKT, $HHNR)
Description	Provides background information about the interview (interviewer ID, date of interview, mode, etc.) for all survey years. It is based on the questionnaire level, hence unique observations are identified only through the combination of SYEAR, PERSNR, PERSNRK and $QSTNR, the identifier of the questionnaire answered. (Note that prior to FiDv3.1, these data were distributed in their yearly format in the datasets ***$paradata***.)
Documentation	Documentation on ***paradatal***

Datasets not in FiD, which are known from the SOEP

There are a couple of datasets which users from the SOEP may miss when skimming through the FiD datasets. Some of the data sets simply could not be produced, because the data are not available, i.e. the information asked for does not include the information needed to create the data sets. This pertains to the datasets ***$bv***, ***bioimmig***, and ***abroad.*** Other datasets could not be provided simply due to the amount of work necessary to produce them (prominently ***$pequiv***). These will be provided as FiD is integrated into the SOEP.

Documentation *ppfad*

Person related meta-dataset

*This documentation is based on the comparable SOEP documentation on **ppfad** and has benefited from previous work. For readability reasons, we do not specifically cite and specify text that has been used directly from the SOEP document.*

General information

This file is designed to support longitudinal analysis when linking personal information from various waves. Each person documented in at least one ***$pbrutto*** file is also in ***ppfad***. The only sorting key is the never-changing person ID number PERSNR. The $HHNR variables contain the current household number of the particular household in which the person lived and was interviewed in at the time of the survey, either as an adult household member or as a child. The $NETTO variables give the wave-specific interview status, and tell which dataset contains a person's record in a particular wave (***$p***, ***$kind***, etc.) or whether the person has permanently left FiD.

In addition, ***ppfad*** contains basic demographic variables, which provide longitudinally consistent information over all waves:

- SEX (sex)
- GEBJAHR (4-digit birth year)
- GEBMONAT (2-digit birth month)
- TODJAHR (4-digit year of death, if applicable)
- IMMIYEAR (year of first immigration to Germany)
- GERMBORN (born in Germany)
- CORIGIN (country of origin)
- MIGBACK (migration background)
- MIGINFO (information source for MIGBACK)

It is recommended that these tested variables be used for cross-sectional and longitudinal analyses. These variables are adjusted on a wave-by-wave basis in the framework of demographic testing.

Further methodological variables (year, four-digit) are included to provide more information on the sample composition:

- EINTRITT: year a person joined FiD, e.g. the year a person was first included in ***$pbrutto***
- AUSTRITT: year a person ultimately left FiD
- ERSTBEFR: year of a person's first interview
- LETZTBEF: year of a person's most recent interview

List of variables

HHNR

Variable label **"Original household number"**
Variable format 7-digit integer

Comment Identifier for the household in which a person lived at the time she was captured in ***$pbrutto*** for the first time. Note that this does not mean that a person had an interview in this household – they could be children or non-responding household members for other reasons as well. People entering a household at a later time receive the original household number as their HHNR, even if the household is a split household.

PERSNR

Variable label **"Never changing Person ID"**
Variable format 8-digit integer

Comment Identifies each person uniquely and for all waves and datasets. Note that some statistical software programs have issues with the length of the variable. E.g. in Stata™, note that when generating a variable identical to PERSNR, you will have to specify that the new variable is in "long" format, i.e. "`gen long varx=persnr`".

$HHNR

Variable label **"Current household number $$$$"**
Variable format 7-digit integer
$ - Wave F10, F11, F12, F13
$$$$ - Year 2010, 2011, 2012, 2013

Comment Identifier for the household in which a person lived at the time of the current wave interview. $HHNR is set to "-2 does not apply" if a person was not in a FiD-household in the respective wave.

PSAMPLE

Variable label **"Subsample"**
Value label PSAMPLE
(61) FiD 2007 Birth Cohort
(62) FiD 2008 Birth Cohort
(63) FiD 2009 Birth Cohort
(64) FiD 2010 Birth Cohort
(65) FiD Screening (sampled 2010)
(66) FiD Screening (sampled 2011)
Variable format 2-digit integer

Comment Note that this variable is included in all datasets, and provides information whether the household or person originates from the cohort or one of the screening samples in FiD. In ***$kind*** it is also named PSAMPLE, in ***hpfad*** it is named HSAMPLE; in all other datasets it is called SAMPLE1, which is analogous to the SOEP notation.

SEX

Variable label **"Sex"**
Value label SEX
(1) male
(2) female
Variable format 1-digit integer

GEBJAHR

Variable label **"Year of birth"**
Variable format 4-digit integer

Comment Respondent's year of birth, checked for consistency across survey years. Note that this information may vary across years due to misreporting, but always provides the best information available. Hence GEBJAHR should be used for any analysis involving the age of respondent (or child).

GEBMONAT

Variable label **"Month of birth"**
Value label GEBMONAT
(1) January
(2) February
(3) March
(4) April
(5) May
(6) June
(7) July
(8) August
(9) September
(10) October
(11) November
(12) December
Variable format 2-digit integer

Comment Respondent's month of birth, checked for consistency across survey years. Note that this information may vary across years due to misinformation, but always provides the best information available.

GEBMOVAL

Variable label **"Month of birth, data source"**
Value label GEBMOVAL
(1) Generated from GEBMONAT (parents)
(2) Info from ***ppfad***
(3) Info from ***$kind***
(4) Info from ***$p***
(5) Info from ***$lela***
(6) Info from ***$bioage***

(7) Info from ***$jugend***

Comment GEBMOVAL provides information from which data source the month of birth was taken from. In principle, GEBMONAT can come from four different sources:

1) it can be taken from a parent-questionnaire;
2) it can come from a person's biography information;
3) it can be provided in a regular person interview;
4) for children, it can come from the household interview and the children's matrix; or
5) it can be generated through other means (e.g. biographical information from other persons).

The hierarchy is exactly in the above order, meaning that the first source is preferred to the second and so on, which also clears conflicts between the sources.

TODJAHR

Variable label **"Year of death, four-digit"**
Variable format 4-digit integer

TODINFO

Variable label **"Year of death, information source"**
Value label TODINFO
(1) from annual survey (pbr_exit)
Variable format 2-digit integer

Comment TODINFO provides information from which data source information on the death of respondent was taken. In FiD, there is only one source at the moment, which is the annual survey.

LOC1989

Variable label **"Where did you live in 1989?"**
Value label LOC1989
(1) East Germany (GDR) incl. East Berlin
(2) West Germany (FRG) incl. West Berlin
(3) abroad (foreign country)
Variable format 2-digit integer

Comment This variable captures where the respondent lived at the time of the German reunification. This information is taken from the second part of the biography questionnaire in FiD, which was fielded in 2011 for the first time. Hence if respondents dropped out of the sample before answering the second part of the biography questionnaire, they will have a missing value in this variable.

$NETTO

Variable label **"Current Wave Survey Status"**

Value label $NETTO

(110) FiD: Person questionnaire
(111) FiD: P-questionnaire & LeLa I
(112) FiD: P-questionnaire & LeLa II
(113) FiD: P-questionnaire & LeLa I+II
(117) FiD: Youth questionnaire, 17
(118) FiD: P-questionnaire and below 17
(119) FiD: P-questionnaire, no household
(120) FiD: Child in part. household (_KIND)
(121) FiD: Parent Questionnaire 1 (0-1)
(122) FiD: Parent Questionnaire 1 2 (1-2)
(123) FiD: Parent Questionnaire 1 3 (2-3)
(124) FiD: Parent Questionnaire 1 4 (5-6)
(125) FiD: Parent Questionnaire 1 5 (7-8)
(126) FiD: Parent Questionnaire 1 6 (9-10)
(130) FiD: in Gross-Sample-HH w/o P-Int.
(131) FiD: Gap Questionnaire (_LUECKE)
(132) FiD: Biography only, no household
(133) FiD: Youth questionnaire, no household
(135) FiD: Parent questionnaire, no household
(160) FiD: Only Questionnaire Without Indiv. And HH Interview
(161) FiD: Gap Interview without HH reference
(162) FiD: Gap Interview with drop out
(170) FiD: Only Participation In Tests, Experiments, etc.
(180) FiD: Individual Without Any Current Information
(181) FiD: Prior Interviewee Without Any Current Information
(188) FiD: Repatriate - (moved abroad before [191])
(189) FiD: Repatriate - (was drop out [190])
(190) FiD: Drop-outs
(191) FiD: Moved abroad
(199) FiD: Deceased

Variable format 3-digit integer
$ - Wave F10, F11, F12, F13

Comment $NETTO follows the SOEP logic in principle, although some of the provided value labels are not relevant for FiD at the moment. The main difference is that FiD uses 3-digit numbers and features an integrated personal interview, capturing the biography information as well as the personal information (codes 111, 112, 113).
Note that code 131 (Gap-Questionnaire) is set retrospectively for the year of the gap, e.g. if a person missed the 2011 interview and fills out the gap questionnaire in 2012, the F11NETTO code is changed accordingly from 130 (gross sample only) to 131.

$NETOLD

Variable label **"Current Wave Survey Status"**
Value label $NETOLD
(0) Person Gap (YPBRUTTO)

(1) Successful Interview (_P [,_JUGEND])
(2) Below Survey Age (_KIND)
(3) Did not participate (_PBRUTTO)
(4) Missing this wave (_PLUECKE)
(5) Interviewee without household interview

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable is a compressed version of the variable $NETTO.

$CASEMAT

Variable label **"Case-match, combined panel households"**
Variable format 6-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable is built after the SOEP to capture the rare event of two FiD-households moving together. This event is extremely unlikely, so this variable is missing ("-2 Does not apply") for all cases at the moment.

$POP

Variable label **"Sample membership $$$$"**
Value label $POP
(1) Private household, German HH-head
(2) Private household, foreign HH-head
(3) Institutional household, German HH-head
(4) Institutional household, foreign HH-head
(5) Not Completed Private HH, German HH-Head
(6) Not Completed Private HH, Foreign HH-Head
(7) Not Completed Institutional HH, German HH-Head
(8) Not Completed Institutional HH, Foreign HH-Head
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13
$$$$ - Year 2010, 2011, 2012, 2013

Comment $POP was derived from ***$pbrutto***, using the variables $pnat and $stell. Note that a value of "-1 No answer" is possible, if the nationality was not reported by the respondent.

$SAMPREG

Variable label **"Current wave sample region $$$$"**
Value label $SAMPREG
(1) West Germany
(2) East Germany
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13
$$$$ - Year 2010, 2011, 2012, 2013

Comment $SAMPREG specifies whether the household is located in the former Western or Eastern part of Germany, the "old" or "new" Länder, respectively. This information uses the old borders as they were defined at the time of the reunification, hence the former West Berlin is labeled as $SAMPREG=1. This information is time-dependent; thus for each year, there is the wave-specific information $SAMPREG.

GERMBORN

Variable label **"Born in Germany or immigration prior to 1949"**
Value label GERMBORN
(1) German born or immigrant prior to 1949
(2) immigrant after 1948
Variable format 1-digit integer

Comment GERMBORN is taken from the biography information the respondents provide. If this information is missing due to item or unit non-response, the variable is set to "-1 No answer".

IMMIYEAR

Variable label **"Year of immigration to Germany after 1948"**
Variable format 4-digit integer

Comment GERMBORN is taken from the biography information the respondents provide. For all individuals born in Germany, this information is set to "-2 Does not apply". If this information is missing due to item or unit non-response, the variable is set to "-1 No answer".

CORIGIN

Variable label **"Country of origin"**
Value label CORIGIN (see below)
Variable format 3-digit integer

(001) Germany	(002) Turkey	(003) Ex-Yugoslavia	(004) Greece
(005) Italy	(006) Spain	(007) Ex-GDR	(010) Austria
(011) France	(012) Benelux	(013) Denmark	(014) Great Britain
(015) Sweden	(016) Norway	(017) Finland	(018) USA
(019) Switzerland	(020) Chile	(021) Rumania	(022) Poland
(023) Korea	(024) Iran	(025) Indonesia	(026) Hungary
(027) Bolivia	(028) Portugal	(029) Bulgaria	(030) Syria
(031) Czech Republic	(032) Russia	(033) Kurdistan	(034) Mexico
(035) Argentina	(036) Cap Verde Is.	(037) Benin	(038) Philippines
(039) Israel	(040) Japan	(041) Australia	(042) India
(043) Afghanistan	(044) Thailand	(045) Jamaica	(046) Saudi-Arabia
(047) Ethiopia	(048) Columbia	(049) Ghana	(050) Bangladesh
(051) Venezuela	(052) Tunisia	(053) Mauritius	(054) Nigeria
(055) Canada	(056) New Zealand	(057) Tanzania	(058) Cyprus
(059) Cuba	(060) Iraq	(061) Brazil	(062) Monaco
(063) Hong Kong	(064) Peru	(065) Sri Lanka	(066) Nepal
(067) Morocco	(068) China	(069) Liechtenstein	(070) Iceland
(071) Ireland	(072) St. Lucia	(073) Moldavia	(074) Kazakhstan
(075) Albania	(076) Lebanon	(077) Kyrgyzstan	(078) Ukraine

(079) Algeria	(080) Mozambique	(081) Egypt	(082) Tajikistan
(083) Vietnam	(084) Somalia	(085) Pakistan	(086) South Africa
(087) UAE	(088) El Salvador	(089) Eritrea	(090) Jordan
(091) Turkmenistan	(092) Costa Rica	(093) Singapore	(094) Burkina Faso
(095) Zambia	(096) Ecuador	(097) Uzbekistan	(098) no nationality
(099) Puerto Rico	(100) Laos	(101) Estonia	(102) Angola
(103) Latvia	(104) Malaysia	(105) Namibia	(106) Montenegro
(107) Belize	(108) Dominican Republic	(109) Nicaragua	(110) Kenya
(111) Libya	(112) Malta	(113) Botswana	(114) Haiti
(115) Trinidad, Tobago	(116) Luxembourg	(117) Belgium	(118) Holland
(119) Croatia	(120) Bosnia-Herzegovina	(121) Macedonia	(122) Slovenia
(123) Slovakia	(124) Paraguay	(125) Guinea	(126) Kuwait
(127) Ivory coast	(128) Malaysia	(129) Samoa	(130) Azerbaijan
(131) Seychelles	(132) Belarus	(133) Uruguay	(134) Bahamas
(135) Uganda	(136) Oman	(137) Micronesia	(138) Mali
(139) Cameroon	(140) Kosovo-Albania	(141) Georgia	(142) Sudan
(143) Congo	(144) Togo	(145) Mongolia	(146) Lithuania
(147) Chad	(148) Armenia	(149) Kurdistan	(150) Liberia
(151) Yemen	(152) Palestine	(153) Free state of Gdansk	(154) Taiwan
(155) Turkmenistan	(156) Africa	(157) Guatemala	(158) Sierra Leone
(159) Panama	(160) East Timor	(161) Bahrain	(162) Senegal
(163) Maldives	(164) Kabarday	(165) Serbia	(166) Gambia
(167) Honduras	(168) Montenegro	(169) Cambodia	(170) Surinam
(171) Guyana	(172) Caucasus	(173) Zimbabwe	(174) Madagascar
(175) Grenada	(222) Eastern Europe	(333) Other, not specified foreign country	

Comment CORIGIN is generated with information from ***$lela*** and the youth questionnaire. If GERMBORN is true (=1), the country is Germany. If children were born before the immigration, the CORIGIN of the mother is transferred to them; if this was not available, the CORIGIN of the father was used.

MIGBACK

Variable label **"Migration Background"**

Value label MIGBACK
- (1) No migration background
- (2) Direct migration background
- (3) Indirect migration background
- (4) Migration background, not differentiated

Variable format 1-digit integer

Comment Generally MIGBACK is defined as in the SOEP. Due to a different information basis there are some slight differences (information about respondent's parents migration status is only available for the second wave respondents, except mother tongue). We use information on country of birth, citizenship, immigration, immigration group, mother tongue of parents and citizenship of parents where available. A direct migration background was assigned if the person migrated herself (i.e. was not born in Germany). Otherwise an indirect migration background was assigned, when one of the person's parents migrated.
For children (under the age of 17), migration background is set to be direct, if their biological parents moved after they were born. Consequently, if the parents had already been in Germany at the time of birth, the migration background becomes indirect. Note that there is a

difference to the SOEP here: as FiD has more detailed information on the parent-child relationship, the migration background is only derived if at least one of the child's biological parents is known.

MIGINFO

Variable label	**"Information Source for MIGBACK"**
Value label	MIGINFO (1) Direct info without parental info (2) Proxy info without parental info (3) Direct info with parental info (4) Proxy info with parental info
Variable format	1-digit integer

Documentation *hpfad*

Household related meta-dataset

Rainer Siegers

*This documentation is based on the comparable SOEP documentation on **hpfad** and has benefited from previous work. For readability reasons, we do not specifically cite and specify text that has been used directly from the SOEP document.*

General information

This dataset is designed to assist linking data from different waves for the substantive evaluation of household level data. This dataset includes all households that ever had contact with FiD in any wave up to the present. Thus a household interview does not necessarily have to have taken place in the current wave; it is enough to simply have appeared once in any one of the "***$hbrutto***" household datasets. The sorting keys are the current household numbers (HHNRAKT and $HHNR). While "old" households retain their old HHNRAKT in case of a move (e.g. F10HHNR=F11HHNR), "new" panel households (e.g. formed by members of an existing panel household that have moved into a new household) receive a different HHNRAKT upon their first interview in a particular wave.

The variables $HHNR also contain the current household number if the household entered the sample in the respective wave. In any years prior to this first interview, the variable $HHNR is "missing", which then helps to prevent mistakes in linking household level data across time. The variables $HNETTO give information on whether it was possible to carry out a HH interview and whether it would pay off to create a link to the corresponding data in that wave`s $H file.

List of variables:

HHNR

Variable label	"**Original household number**"
Variable format	7-digit integer
Comment	HHNR denotes the original household number of the household and allows relating households that have split over the years. For a household in its first wave, HHNR always equals HHNRAKT and $HHNR. After that, a HHNR differing from the current household number indicates a break-off household that has successfully been followed. Note that individuals moving into a break-off household receive the HHNR of the household, even though they never lived in it.

HHNRAKT

Variable label	"**Current** h**ousehold number**"
Variable format	7-digit integer
Comment	HHNRAKT is the current household number. It is identical to the wave specific household number, $HHNR. HHNRAKT is not updated if a household is no longer in the sample, hence the variable is never missing (i.e. "-2 does not apply"). This is a key difference to $HHNR, which can be missing (either if a household is not yet in the panel in a specific wave or has dropped out).

HSAMPLE

Variable label	"**Subsample**"
Value label	HSAMPLE (61) FiD 2007 Birth Cohort (62) FiD 2008 Birth Cohort (63) FiD 2009 Birth Cohort (64) FiD 2010 Birth Cohort (65) FiD Screening (sampled 2010) (66) FiD Screening (sampled 2011)
Variable format	2-digit integer
Comment	HSAMPLE is fixed across waves and households and denotes the sample membership at first contact with FiD (at the point of selection of the original household into the sample).

$HHNR

Variable label	"**Household number $$$$**"
$ - Wave	F10, F11, F12, F13
$$$$ - year	2010, 2011, 2012, 2013
Variable format	7-digit integer
Comment	$HHNR denotes the household's current identifier. It is set to missing ("-2 does not apply") in case the household is not yet or no longer in the

sample. E.g., all households in the Screening-Sample of 2011 (HSAMPLE=66) have a F10HHNR of "-2", as they entered the sample in 2011 only.

$SAMPREG

Variable label	"**Current wave sample reason $$$$"**
Value label	$SAMPREG (1) West Germany (former FRG) (2) East Germany (former GDR)
$ - Wave	F10, F11, F12, F13
$$$$ - year	2010, 2011, 2012, 2013
Variable format	1-digit integer
Comment	$SAMPREG specifies whether the household is located in the former Western or Eastern part of Germany, the "old" or "new" Länder, respectively. This information uses the old borders as they were defined at the time of the reunification, hence the former West Berlin is labeled as $SAMPREG=1. This information is time-dependent; thus for each year, there is the wave-specific information $SAMPREG.

$HPOP

Variable label	"**Sample membership $$$$**"
Value label	$HPOP (1) private household, German head (2) private household, foreign head (3) institutional household, German head (4) institutional household, foreign head
$ - Wave	F10, F11, F12, F13
$$$$ - year	2010, 2011, 2012, 2013
Variable format	1-digit integer
Comment	$HPOP is derived from $WUM2 (in ***$hbrutto*** differentiating private from institutional households) as well as $PNAT and $STELL (nationality and relationship to household head in ***$hbrutto***). Missing values are imputed taking into account respondent's history. Thus the only admissible missing value is "-2 does not apply".

$HNETTO

Variable label	"**Wave survey status $$$$**"
Value label	$HNETTO (1) successful household interview (2) household in gross sample only
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
$$$$ - year	2010, 2011, 2012, 2013
Comment	The wave-specific survey status sets $HNETTO to 1 if there is an interview in the respective year, i.e. an entry in ***$h*** or in the ***$p*** (note

that for very few cases there is no household interview, but one or more person interviews). For $HNETTO=2 there is only an entry in the file ***$hbrutto***, but the household was not interviewed (the reason for which is given by the variable $HERGS in the file ***$hbrutto***).

SINGPA

Variable label	"**Single-parent-household at first contact**"
Value label	SINGPA (1) Yes (2) No
Variable format	1-digit integer
Comment	SINGPA is a time-fixed indicator providing information whether a household was defined as "single-parent-household" the first time it was interviewed. This information belongs to the original household and is thus associated with the variable HHNR. It is carried over to all split households from the original household. In the Screening-Samples 2010 and 2011, this corresponds to the sampling-design-definition (exactly one adult person with at least one child younger than 18) at the point of selection of the original household into the sample. For the Cohort-Sample, this information was not used during sampling, but during the process of generating the integrated weights.

LRGFAM

Variable label	"**Large-family-household at first contact**"
Value label	LRGFAM (1) Yes (2) No
Variable format	1-digit integer
Comment	LRGFAM is a time-fixed indicator providing information whether a household was defined as "large-family-household" the first time it was interviewed. This information belongs to the original household and is thus associated with the variable HHNR. It is carried over to all split households from the original household. In the Screening-Samples 2010 and 2011, this corresponds to the sampling-design-definition (at least three children younger than 18 years in household) at the point of selection of the original household into the sample. For the Cohort-Sample, this information was not used during sampling, but during the process of generating integrated weights.

LOWINC

Variable label	"**Low-income-household at first contact**"
Value label	LOWINC (1) Yes (2) No (3) Unknown
Variable format	1-digit integer

Comment

LOWINC is a time-fixed indicator providing information whether a household was defined as “low-income-household” the first time it was interviewed. This information belongs to the original household and is thus associated with the variable HHNR. It is carried over to all split households from the original household. In the Screening-Sample 2010, this corresponds to the sampling-design-definition at the point of selection of the original household into the sample. For the cohort sample, this information was not used during sampling, but during the process of generating integrated weights. LOWINC is set to “-2 Does not apply” for all households in the Screening-Sample 2011, as this information was not used for sampling or weighting purposes.

The value “(3) Unknown” indicates that no low-income-decision could be made due to missing values on the net household income in 2010 (F10H045A). Given a valid value on net household income, a low-income-household is defined as

a) a household with one adult and children and a maximum net income of 1500€,
b) a household with more than one adult and one child and a maximum net income of 2000€ or
c) a household with more than one adult and more than one child and a maximum net income of 2500€ a month.

If possible, the information of F10H045B is also used.

Documentation *$pgen*

Person Related Status Variables and Generated Variables

Stefan Damerow, Mathis Fräßdorf (geb. Schröder) and Juliana Werneburg

This documentation is based on the comparable SOEP documentation on PGEN and has benefited from previous work of Silke Anger, Joachim Frick, Markus Grabka, Jan Goebel, Peter Krause, Henning Lohmann, Olaf Groh-Samberg, and Pia Schober. For readability reasons we do not specifically cite and specify text that has been used directly from the SOEP document.

General information

The ***$pgen*** datasets provide individual level information for each survey wave. Similar to the data included in ***$hgen*** for the household level, the ***$pgen*** files give the researcher an easier access to individual information over the waves. Variable names and formats are consistent over all years included, such that a combination of these files facilitates a comparison of different years. In addition, some variables are completely generated or imputed, and hence provide information which is not included in the regular questionnaire. The following documentation lists all variables in ***$pgen*** and provides information on how they were generated. For all variables, the information provided looks as follows:

Variable label	Provides the label of the variable as it is given in the dataset. Variables are given in CAPTIAL letters, even though they might appear in small letters in the dataset. This is simply for readability.

Value labels	*LBLNME* In case *VARNME* is categorical, *LBLNME* specifies the labels for each category, and the value labels are listed here. Note that the standard missing value labels (-1: No answer; -2: Does not apply; -3: Not valid) are not listed, but apply to all variables in this dataset.
Variable format	Specifies the format for each variable, e.g. “1-digit integer” or “string”.
$$ - Survey Years	Specifies the years for which the variable is provided. This is provided for 2000+, such that “10” refers to 2010, etc.
Comment:	Provides more detailed information on the generating process, also on the population the variable is specified for, if necessary. Here, variables used, changes between waves, or any other anomalies are mentioned and their relevance explained.

Some variables will be written forward from previous waves, as they are only collected once and do not necessarily change every year. These variables include some employment variables, the education variables, marital status, and partner variables.

If you have questions regarding ***$pgen*** data for the FiD-distribution, unless noted otherwise, please contact Mathis Schröder at +49 (0)30 / 89789 - 222.

List of variables:

EMPLST$$

Variable label	**"Employment status"**
Value label	EMPLST$$ (1) full-time employment (2) regular part-time employment (3) vocational training (4) marginal, irregular part-time employment (5) not employed (6) sheltered workshop
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable is generated from the annual question on current labor market participation, which has a central filter function in the questionnaire to separate employed from non-employed respondents for further questions. It is designed to provide consistent longitudinal data on employment status across all waves.
The category "not employed" comprises non-working individuals, those in military/community service, those on maternity leave, and employed persons in a phased retirement scheme (*Altersteilzeit*), whose current actual working hours are zero. The additional category "sheltered workshop" is included for disabled persons in sheltered employment.

EMPLST$$ is supplemented by the variable LFS$$, which differentiates among persons who are not employed.

LFS$$

Variable label	**"Labor force status"**
Value label	LFS$$ (1) non-working without further information (2) non-working, and older than 65 (3) non-working, in training program (4) non-working, on maternity leave (5) non-working, in military/community service (6) non-working, and registered unemployed (8) non-working, but sometimes second job (9) non-working, but working past 7 days (10) non-working, but regular second job (11) working (12) working, but non-working past 7 days
Variable format	2-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable is based on the annual question on current labor market participation, combined with additional information on activities of non-working individuals. The number of values assigned has been based on a large number of highly differentiated answer categories.
LFS$$ provides a differentiation between "working" (Code 11-12) and "non-working" (Code 1-10). Non-employment is split further in order

to make it possible to efficiently apply different labor market concepts in studying the data. To construct this variable, the variables on employment status, age, maternity leave, second jobs, registration at the employment office, participation in paid work during the past 7 days and training status are used.

For respondents who have multiple status codes and different values for this variable, the following hierarchy was used to determine which of the values would play the determining role (increasing dominance)

11 - working
1 - non-working without further information
2 - non-working, and older than 65
3 - non-working, and currently in a training program
6 - non-working, and registered unemployed
4 - non-working, on maternity leave
5 - non-working, in military/community service
9 - non-working, but working past 7 days
10 - non-working, but regular second job
8 - non-working, but occasional second job
12 – working, but non-working past 7 days

LFS$$ supplements the variable EMPLST$$, which differentiates among persons who are employed.

JOBCH$$

Variable label	**"Job change"**
Value label	JOBCH$$ (1) not employed (2) employed, no change (3) employed, no info if change (4) employed, with change (5) first time employed
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable indicates a change of job since the last interview for respondents with a follow-up interview, whereas for first-time respondents, the information refers to a change of job since the beginning of the previous year. JOCH$$ is generated based on the central filter variable, which indicates whether a respondent has changed jobs since the previous year.

The variable is used to integrate the information for first-time respondents and follow-up respondents. The following hence only applies from the second wave (2011) onwards, where both types of respondents appear for the first time.

The variable also identifies respondents who have entered employment for the first time. The variable will provide consistent longitudinal information on job changes as well. The JOBCH$$ variable is

generated by correcting the original job change information in various ways:
1. We check whether the job changes stated by a respondent in two consecutives interviews refer to one and the same job change. The date of the job change and the interview month are used to correct double entries.
2. If the respondent indicates a job change with a date before the previous interview but did not state a job change in the previous interview, this is coded as a job change in the current interview.
3. If a respondent indicates no job change and was not employed in the previous interview, this is coded as "no job change" because there could have been short-term employment spells between the previous year's and this year's interview.
4. Respondents can be "first-time employed" only once. If a respondent states being "first-time employed" for a second time, this is coded as "employed, with change".

EXPFT$$

Variable label	"**Working experience full-time employment**"
Variable format	3-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

Full-time working experience

This variable reflects the total length of full-time employment in the respondent's career up to the point of the interview. The variable is created by combining monthly information from the calendar dataset ***artkalen*** (which provides monthly information on activity status since an individual entered FiD) and annual information from the biographical dataset ***pbiospe*** (which provides information on activity status over the individual's life course). EXPFT$$ gives the length of time in months – different from the SOEP, where it is provided in years with months in decimal form.

If there is no monthly calendar data available in a given year of a respondent's career, the annual data from ***pbiospe*** is used for that year. If the year in which a spell started and ended is the same, and if there is no monthly data, a spell of 6 months is assumed. Persons without annual data (not contained in ***pbiospe***) are only assigned a non-missing value for this variable if they joined FiD by the age of 18 and if there is calendar data for them in ***artkalen***.

Persons whose life course has been observed completely but with no spell of full-time employment are assigned the code (0). The code ("-1 No answer") is assigned to all persons whose life course has not been observed completely. Persons with inconsistent information receive a ("-3 Answer implausible").

Note that differently to the SOEP, respondents in FiD receive a value of "-2 Does not apply" if their biography data has not been collected yet.

For example, this is the case for individuals, who only participate in 2010. For a person to have valid values in EXPFT$$, the respective biography part has to be answered ($LELTYP=2 or 3).

Please also see EXPPT$$ and EXPUE$$.

EXPPT$$

Variable label	"**Working experience part-time employment**"
Variable format	3-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

Part-time working experience

This variable reflects the total length of part-time employment in the respondent's career up to the point of the interview. The variable is created by combining monthly information from the calendar dataset ***artkalen*** (which provides monthly information on activity status since an individual entered FiD) and annual information from the biographical dataset ***pbiospe*** (which provides information on activity status over the individual's life course). EXPFT$$ gives the length of time in months – different from the SOEP, where it is provided in years with months in decimal form.

If there is no monthly calendar data available in a given year of a respondent's career, the annual data from ***pbiospe*** is used for that year. If the year in which a spell started and ended is the same, and if there is no monthly data, a spell of 6 months is assumed. Persons without annual data (not contained in ***pbiospe***) are only assigned a non-missing value for this variable if they joined FiD by the age of 18 and if there is calendar data on them in ***artkalen***.

Persons whose life course has been observed completely but with no spell of part-time employment are assigned the code (0). The code ("-1 No answer") is assigned to all persons whose life course has not been observed completely. Persons with inconsistent information receive a ("-3 Answer implausible").

Note that differently to the SOEP, respondents in FiD receive a value of "-2 Does not apply" if their biography data has not been collected yet. For example, this is the case for individuals, who only participate in 2010. For a person to have valid values in EXPPT$$, the respective biography part has to be answered ($LELTYP=2 or 3).

Please also see EXPFT$$ and EXPUE$$.

EXPUE$$

Variable label	"**Unemployment experience**"
Variable format	3-digit integer

$$ - Survey Years $$=10, 11, 12, 13

Comment Unemployment experience

This variable reflects the total length of unemployment in the respondent's career up to the point of the interview. The variable is created by combining monthly information from the calendar dataset ***artkalen*** (which provides monthly information on activity status since an individual entered FiD) and annual information from the biographical dataset ***pbiospe*** (which provides information on activity status over the individual's life course). EXPFT$$ gives the length of time in months – different from the SOEP, where it is provided in years with months in decimal form.

If there is no monthly calendar data available in a given year of a respondent's career, the annual data from ***pbiospe*** is used for that year. If the year in which a spell started and ended is the same, and if there is no monthly data, a spell of 6 months is assumed. Persons without annual data (not contained in ***pbiospe***) are only assigned a non-missing value for this variable if they joined FiD by the age of 18 and if there is calendar data on them in ***artkalen***.

Persons whose life course has been observed completely but with no spell of unemployment are assigned the code (0). The code ("-1 No answer") is assigned to all persons whose life course has not been observed completely. Persons with inconsistent information receive a ("-3 Answer implausible").

Note that differently to the SOEP, respondents in FiD receive a value of "-2 Does not apply" if their biography data has not been collected yet. For example, this is the case for individuals, who only participate in 2010. For a person to have valid values in EXPUE$$, the respective biography part has to be answered ($LELTYP=2 or 3).

Please also see EXPFT$$ and EXPPT$$.

TENURE$$

Variable label **"Tenure in months"**
Variable format 3-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment The variable TENURE$$ is designed to offer data on the length of time with the firm at the point in time of the interview for all employed persons. This variable is generated from the respondent's start date with the current employer and the start date of the current position if there was a job change.

The variable provides consistent longitudinal information on the length of time with the same employer. Data that show longitudinal inconsistencies are corrected. In case of no job change, the information on the start date with the current employer given in the earliest

interview available is treated as dominant and carried forward to the subsequent years. In case of a job change, the information on the start of the current position is used and carried forward to the subsequent years. In the case that a respondent starts working again after a period of non-employment, he/she is assumed to have returned to the former employer if the start date with the current employer was before the previous interview date. In this case, the start date with the current employer given in the previous interview is treated as dominant. Otherwise, the present information on the start date with the current employer is used and carried forward to the subsequent years. For respondents who are assumed to have returned to their former employer, the full length of time with the firm is calculated. There is no deduction for the time during which the respondent was not employed. Both monthly and annual information is used in the variable, which is provided – differently to the SOEP – in months.

JOBTRN$$

Variable label	**"Working in occupation trained for"**
Value label	JOBTRN$$ (1) yes (2) no (3) currently in training (4) has no job training
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	This variable is designed to offer annual data on all employed persons, indicating whether they are working in the occupation they were trained for. This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

REQUTR$$

Variable label	**"Required job training"**
Value label	REQUTR$$ (1) no training (2) brief on-the-job training (3) extensive on-the-job training (4) attended courses (5) completed vocational training (8) Fachhochschule degree (9) University degree
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	This variable is designed to provide annual data on required job training for all employed persons. The variable is generated using questions on required formal education and required on-the-job-training which are categorized into up to seven independent variables with 0/1 coding. Out

of these, the highest available level of required training is used for the generation of the status variable.
This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.
The missing value ("-2 Does not apply") was assigned to all non-employed persons and also includes persons in occupational training, in occupational retraining programs, and those doing an internship at the time of the survey.

COSIZE$$

Variable label **"Size of company"**
Value label COSIZE$$
(1) less than 5
(2) 5 to 10
(3) 11 to 20
(4) [up to 1990 less than 20]
(5) [1991-2004 5 to 20]
(6) 20 to 100
(7) 100 to 200
(8) [up to 1998 20 to 200]
(9) 200 to 2000
(10) 2000 or more
(11) Self-employed without other employees
Variable format 2-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment This variable is designed to offer annual data on company size for all employed persons. The codes 4, 5, and 8 were given in the data to make this variable comparable to the SOEP logic, although they will not be set in FiD. (These codes were necessary due to the differentiation of items for small and medium-sized companies over the years.)
This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.
Please also see COSIZE$$ for a broader categorization of the firm size, again comparable to the SOEP logic.
(Note that COSIZE$$ was named FSIZE$$ in FiD v1.2. To avoid confusions with the identically named variable in ***$hgen***, we renamed the variable in ***$pgen***. Nothing else has changed.)

CRSIZE$$

Variable label **"Core size category of the company"**
Value label CRSIZE
(1) fewer than 20
(2) 20 to 200
(3) 200 to 2000
(4) 2000 or more
(5) Self-employed without other employees

Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment This variable is designed to provide annual data on the core size category of the company for all employed persons compared to the SOEP logic.
This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.
(Note that CRSIZE$$ was named CFSIZE$$ in FiD v1.2.)

CIVILS$$

Variable label **"Civil Service"**
Value label CIVILS$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment This status variable is designed to provide annual data on employment in the civil service for all employed persons. This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

OCCPOS$$

Variable label **"Occupational Position"**
Value label OCCPOS$$ (see below)
Variable format 3-digit integer
$$ - Survey Years $$=10, 11, 12, 13

(-1) No Answer
(10) Not Employed
(11) In Education
(12) Unemployed, Not Employer
(13) Pensioner
(15) Military, Community Service
(110) Apprentice
(120) Apprentice, Trainee Industry Technology
(130) Apprentice, Trainee Trade and Commerce
(140) Trainee, Intern
(150) Research assistant
(210) Untrained Worker
(220) Semi-Trained Worker
(230) Trained Worker
(240) Foreman, Team Leader
(250) Foreman
(310) Agricultural Worker
(320) Agricultural Specialist
(330) Agricultural Foreman
(340) Agricultural Manager
(410) Self-Employed Farmer
(411) Self-Employed Farmer, No Employees
(412) Self-Employed Farmer LE 9 Employees
(413) Self-Employed Farmer GT 9 Employees
(420) Free-Lance Professional
(421) Free-Lance Professional, No Employees
(422) Free-Lance Professional, LE 9 Employees

(423) Free-Lance Professional, GT 9 Employees
(430) Other Self-Employed No Or LE 9 Employees
(431) Other Self-Employed No Employees
(432) Other Self-Employed LE 9 Employees
(433) Other Self-Employed GT 9 Employees
(440) Help In Family Business
(510) Foreman
(520) Employee with Simple Tasks
(520) W-Collar Worker with Simple Tasks
(521) Untrained Employee with Simple Tasks
(521) Untrained W-Collar Worker with Simple Tasks
(522) Trained Employee with Simple Tasks
(522) Trained W-Collar Worker with Simple Tasks
(530) Qualified Professional
(540) H. Qualified Professional
(550) Managerial
(610) Low-Level Civil Service
(620) Middle-Level Civil Service
(630) High-Level Civil Service
(640) Executive Civil Service

Comment

The variable represents a compilation of all relevant information on current occupational position. It is generated by combining information on "occupational group", "unemployed (yes/no)", "military/community service", "in education (yes/no)" and "pensioner". A hierarchical scheme is used to determine which data is given precedence when a variety of divergent information exists (increasing dominance):

10 – not employed
13 – pensioner
11 – currently in education
15 – military / community service
12 – registered unemployed
110-150 – apprentice
410-440 – self-employed
210-250 – manual laborer
510-550 – employee
610-640 – civil service

The categories (150) and (310) to (340) were only assigned in the SOEP to respondents in East Germany in 1990. In OCCPOS$$, non-working persons are only assigned to the category (13) "pensioner" if they are recipients of retirement pension or if they are recipients of widow's pension AND are older than 60 years. Moreover, if there is missing information on pension receipt, additional information from ***artkalen*** (retrospective information from the activity calendar for the previous year) will be used in the subsequent waves in the generation process to determine whether a person was in retirement or early retirement (*Vorruhestand*) at the time of the interview.

CLASS$$

Variable label **"StaBuA 1992 Job Classification"**
Value label CLASS$$ (see below)
Variable format 4-digit integer

$$ - Survey Years $$=10, 11, 12, 13

Comment

This variable is designed to provide annual data on job classification for all employed persons according to the classification of the German Federal Statistical Office (StaBuA). Respondents answer the question on their current occupational title in their own words, and this response is entered into a blank in the questionnaire. Due to data protection regulations, this information cannot be provided to data users and was therefore completely recoded by Infratest Sozialforschung. This recoding has been documented in Hartmann/Schütz 2002.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

The occupational classification of the German Federal Statistical Office differentiates among six main occupational types (see next page).

I	KLAS-Codes 0100-0629	**Berufe in der Land-, Tier-, Forstwirtschaft und im Gartenbau**
II	KLAS-Codes 0700-0809	**Bergleute, Mineralgewinner**
III		**Fertigungsberufe**
IIIa	KLAS-Codes 1000-1129	Berufe in der Steinbearbeitung und Baustoffherstellung
IIIb	KLAS-Codes 1200-1359	Keramik-, Glasberufe
IIIc	KLAS-Codes 1400-1539	Chemie-, Kunststoffberufe
IIId	KLAS-Codes 1600-1799	Berufe in der Papierherstellung, -verarbeitung und im Druck
IIIe	KLAS-Codes 1800-1859	Berufe in der Holzverarbeitung, Holz- und Flechtwarenherstellung
IIIf	KLAS-Codes 1900-2459	Berufe in der Metallerzeugung und –bearbeitung
IIIg	KLAS-Codes 2500-3099	Metall-, Maschinenbau- und verwandte Berufe
IIIh	KLAS-Codes 3100-3189	Elektroberufe
IIIi	KLAS-Codes 3200-3239	MontiererInnen und Metallberufe, a.n.g.
IIIk	KLAS-Codes 3300-3619	Textil- und Bekleidungsberufe
IIIl	KLAS-Codes 3700-3789	Berufe in der Lederherstellung, Leder- und Fellverarbeitung
IIIm	KLAS-Codes 3900-4359	Ernährungsberufe
IIIn	KLAS-Codes 4400-4729	Hoch-, Tiefbauberufe
IIIo	KLAS-Codes 4800-4929	Ausbauberufe, PolsterInnen
IIIp	KLAS-Codes 5000-5069	Berufe in der Holz- und Kunststiffverarbeitung
IIIq	KLAS-Codes 5100-5149	MalerInnen, LackiererInnen und verwandte Berufe
IIIr	KLAS-Codes 5200-5239	WarenprüferInnen, VersandfertigmacherInnen
IIIs	KLAS-Codes 5300-5319	HilfsarbeiterInnen ohne nähere Tätigkeitsangabe
IIIt	KLAS-Codes 5400-5509	MaschinistInnen und zugehörige Berufe
IV		**Technische Berufe**
IVa	KLAS-Codes 6000-6129	IngenieurInnen, ChemikerInnen, PhysikerInnen, MathematikerInnen
IVb	KLAS-Codes 6200-6529	TechnikerInnen, Technische Sonderfachkräfte
V		**Dienstleistungsberufe**
Va	KLAS-Codes 6600-6899	Warenkaufleute
Vb	KLAS-Codes 6900-7069	Dienstleistungskaufleute und zugehörige Berufe
Vc	KLAS-Codes 7100-7449	Verkehrsberufe
Vd	KLAS-Codes 7500-7899	Organisations-, Verwaltungs-, Büroberufe
Ve	KLAS-Codes 7900-8149	Ordnungs- und Sicherheitsberufe
Vf	KLAS-Codes 8200-8399	Schriftwerkschaffende, -ordnende und künstlerische berufe
Vg	KLAS-Codes 8400-8599	Gesundheitsdienstberufe
Vh	KLAS-Codes 8600-8949	Sozial- und Erziehungsberufe, anderweitig nicht genannte geistes- und sozialwissenschaftliche Berufe
Vi	KLAS-Codes 9000-9379	Sonstige Dienstleistungsberufe
VI	KLAS-Codes 9700-9979	**Sonstige Arbeitskräfte**

Because of gaps in the answers provided by respondents, the following "new" codes were created:

9711	Mithelfende Familienangehörige außerhalb der Landwirtschaft, anderweitig nicht genannt
9811	Auszubildende mit (noch) nicht feststehendem Ausbildungsberuf
9821	Praktikanten/Praktikantinnen, Volontäre/ Volontärinnen mit (noch) nicht feststehendem Beruf
9911	Facharbeiter/innen, ohne nähere Tätigkeitsangabe
9921	Heimarbeiter/innen, ohne nähere Tätigkeitsangabe
9931	Vorarbeiter/innen, Gruppenleiter/innen, ohne nähere Tätigkeitsangabe
9971	Sonstige Arbeitskräfte, ohne nähere Tätigkeitsangabe

Detailed description: Statistisches Bundesamt (1996) Bevölkerung und Erwerbstätigkeit, Fachserie 1, Reihe 4.1.2., Beruf, Ausbildung und Arbeitsbedingung der Erwerbstätigen 1995 (Ergebnisse des Mikrozensus). Stuttgart Metzler-Poeschel. pp. 317-323. Hartmann/Schütz (2002) Die Klassifikation der Berufe und der Wirtschaftzweige im Sozio-oekonomischen Panel – Neuvercodung der Daten 1984 – 2001. Infratest Sozialforschung, München.

IS88$$

Variable label **"4-digit ISCO-88 Occupation Code"**
Value label ISCO$$ (see below)
Variable format 4-digit integer
$$ - Survey Years $$=10, 11, 12, 13

(1000) Legislators, senior officials, and managers
(1100) Legislators and senior officials
(1200) Corporate managers
(1300) Managers of small enterprises
(2000) Professionals
(2100) Physical, mathematical, and engineering science professionals
(2200) Life science and health professionals
(2300) Teaching professionals
(2400) Other professionals
(3000) Technicians and associate professionals
(3100) Physical and engineering science associate professionals
(3200) Life science and health associate professionals
(3300) Teaching associate professionals
(3400) Other associate professionals
(4000) Clerks
(4100) Office clerks
(4200) Customer services clerks
(5000) Service Workers and shop and market sales workers
(5100) Personal and protective services workers
(5200) Models, salespersons, and demonstrators
(6000) Skilled agricultural and fishery Workers
(6100) Skilled agricultural and fishery workers
(7000) Craft and related trades workers
(7100) Extraction and building trades workers
(7200) Metal, machinery, and related trades workers
(7300) Precision, handicraft, craft printing and related trades workers
(7400) Other craft and related trades workers
(8000) Plant and machine operators and assemblers
(8100) Stationary plant and related operators
(8200) Machine operators and assemblers
(8300) Drivers and mobile plant operators
(9000) Elementary occupations
(9100) Sales and services elementary occupations
(9200) Agricultural, fishery, and related laborers
(9300) Laborers in mining, construction, manufacturing, and transport

Comment This variable is designed to provide annual data on occupational activity for all employed persons according to the International Standard Classification of Occupations ISCO-88. Respondents answer the question on their current occupational title in their own words, and this response is entered into a blank in the questionnaire. Due to data protection regulations, this information cannot be provided to data users and was therefore completely recoded by Infratest Sozialforschung. This recoding has been documented in Hartmann/Schütz 2002.

ISCO-88 is a strictly four-digit classification, and this variable is therefore coded in four-digit form. In contrast to the previous version of the classification system, ISCO-68, ISCO-88 does not use blanks if there is not adequate information for specific coding, but uses zeros instead. Thus 4000 stands for an unspecified office job; 2300 stands for teachers and 2000 stands for scientists, both without closer specification. There is no conversion key since the two classifications differ significantly. The SOEP data distribution 1984-2001 replaced ALL earlier data distributions with the ISCO-88-coding.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Detailed description: International Labour Office (ILO) (1990) ISCO-88; International Standard Classificaton of Occupation, Genf.
Hartmann/Schütz (2002) Die Klassifikation der Berufe und der Wirtschaftzweige im Sozio-oekonomischen Panel – Neuvercodung der Daten 1984 – 2001. Infratest Sozialforschung, München.

NACE$$

Variable label **"Two-digit NACE Industry – Sector"**
Value label NACE$$ (1-100) (see below)
Variable format 3-digit integer
$$ - Survey Years $$=10, 11, 12, 13

(1) Agriculture, Hunting, Related Service Activities
(2) Forestry, Logging, Related Service activities
(5) Fishing, Operation Of Fish Hatcheries And Fish Farms
(10) Mining Of Coal And Lignite; Extraction Of Peat
(11) Extraction Of Crude Petroleum And Natural Gas
(12) Mining Of Uranium And Thorium Ores
(13) Mining Of Metal Ores
(14) Other Mining And Quarrying
(15) Manuf. Food Products And Beverages
(16) Manuf. Tobacco Products
(17) Manuf. Textiles
(18) Manuf. Wearing Apparel; Dressing And Dyeing Of Fur
(19) Tanning, Dressing Of Leather; Manuf. luggage, Footwear
(20) Manuf. Wood Products, Except Furniture
(21) Manuf. Pulp, Paper And Paper Products
(22) Publishing, Printing And Reproduction Of Recorded Media
(23) Manuf. Coke, Refined Petroleum Prod, Nuclear Fuel
(24) Manuf. Chemicals And Chemical Products
(25) Manuf. Rubber And Plastic Products
(26) Manuf. Other Non-metallic Mineral Products
(27) Manuf. Basic Metals
(28) Manuf. Fabricated Metal Prod., Ex. Machinery And Equip
(29) Manuf. Machinery And Equipment NEC
(30) Manuf. Office Machinery And Computers
(31) Manuf. Electrical Machinery And Apparatus NEC
(32) Manuf. Radio, Television And Communication Equipment
(33) Manuf. Medical, Precision And Optical Instruments
(34) Manuf. Motor Vehicles, Trailers And Semi-trailers
(35) Manuf. Other Transport Equipment
(36) Manuf. Furniture; Manufacturing NEC
(37) Recycling
(40) Electricity, Gas, Steam And Hot Water Supply
(41) Collection, Purification And Distribution Of Water
(45) Construction
(50) Sale, Maint., Repair Motor Vehicles; Retail Car Gas
(51) Wholesale Trade, Commission Trade, Ex. Motor Vehicles
(52) Retail, Ex. Motor vehicles, Motorcycles; Repair
(55) Hotels And Restaurants
(60) Land Transport; Transport Via Pipelines
(61) Water Transport
(62) Air Transport
(63) Supporting, Aux. Transport Activities; Travel agencies
(64) Post And Telecommunications
(65) Financial Intermediation, Ex. Insurance, Pension Funding
(66) Insurance And Pension Funding, Ex. Compulsory SocSec

(67) Activities Auxiliary To Financial Intermediation
(70) Real Estate, Property Activities
(71) Renting Of Machinery, Equip Wo. Oper., Pers,HH Goods
(72) Computer And Related Activities
(73) Research And Development
(74) Other Business Activities
(75) Public Administration And Defense; Compulsory SocSec
(80) Education
(85) Health And Social Work
(90) Sewage And Refuse Disposal, Sanitation And Related
(91) Activities Of Membership Organizations NEC.
(92) Recreational, Cultural And Sporting Activities
(93) Other Service Activities
(95) Private Households With Employed Persons
(96) Industry - NEC
(97) Handcraft, Trade - NEC
(98) Services - NEC
(99) Extra-territorial Organizations And Bodies
(100) Manufacturing - NEC

Comment

This variable is designed to provide annual data on the industry of economic activity for all employed persons according to the Statistical Classification of Economic Activities in the European Community (Nomenclature des statistiques des activités économiques de la Communauté européenne - NACE). Respondents answer the question in their own words regarding the industry in which they are currently working, and this response is entered into a blank in the questionnaire.
In order to facilitate international comparability, the European industry standard classification system is used by Infratest Sozialforschung to recode this information. This recoding has been documented in Hartmann/Schütz 2002.

The codes in NACE Rev.1 also correspond to ISIC Rev.3 (International Standard Classification of All Economic Activities). Please note that special codes 96-98 as well as 100 were assigned by Infratest Sozialforschung whenever respondents did not provide a more detailed answer.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Detailed description: Hartmann/Schütz (2002) Die Klassifikation der Berufe und der Wirtschaftzweige im Sozio-oekonomischen Panel – Neuvercodung der Daten 1984 – 2001. Infratest Sozialforschung, München.

ISEI$$

Variable label	**"ISEI-status 88 by Ganzeboom (based on ISCO88)"**
Variable format	2-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable reflects the Standard International Socio-Economic Index of Occupational Status for all employed persons. The ISEI Index was developed in 1992 by Ganzeboom, De Graaf, Treiman, and De Leew based on information about income, education, and occupation. Technically, ISEI was created by scaling the ISCO88 classification. The values for the variable range between 16 and 90. In contrast to the prestige scores of Ganzeboom and Treiman (1996) and Wegener (1988), ISEI is a measure of socio-economic status.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Please also see occupational prestige scores (SIOPS$$, MPS$$) and occupational class (EGP$$).

Detailed description: Ganzeboom, H. B. G. / De Graaf, P .M. / Treiman, D. J. / De Leew, J.(1992) A Standard International Socio-Economic Index of Occupation Status, In Social Science Research 21 1-56

EGP$$

Variable label	**"Erikson and Goldthorpe Class Category"**
Value label	EGP$$

(1) high service
(2) low service
(3) routine non-manual
(4) routine service-sales
(5) self-employed with employees
(6) self-employed without employees
(8) skilled manual
(9) semi-unskilled manual
(10) farm labor
(11) self-employed farmer
(15) not working – unemployed
(18) not working – pensioner

Variable format	2-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable gives the occupational class for all employed persons. EGP$$ is derived from the Standard International Socio-Economic Index of Occupational Status (ISEI). Technically, the variable was created by scaling the ISCO-88 classification. In addition, it is based on

information about income, education and occupation. The EGP Index was documented by Ganzeboom/Treiman in 1996 and revised in 2003.

The values for the variable range between 1 and 11; additional categories are (15) not working – registered unemployed and (18) not working – pensioner.

Non-working persons are only assigned to the category "not working – pensioner" if they are recipients of retirement pension or if they are recipients of widow's pension AND are older than 60 years. Moreover, if there is missing information on pension receipt, additional information from ***artkalen*** (retrospective information from the activity calendar for the previous year) will be used in future waves in the generation process to determine if a person was in retirement or early retirement (*Vorruhestand*) at the time of the interview. Hence, the category "not working – pensioner" in the most recent wave will be updated with retrospective information of the following wave. All other non-working persons are assigned to category (-2) "does not apply" as long as they are not registered as unemployed (category 15).

Annual information on the occupational position is used to generate the EGP-categories for the self-employed. In case no information on the number of employees is available, the EGP$$-categories (5) and (6) contain information on the firm size for self-employed persons.

Based on the new classification developed by Ganzeboom/Treiman (2003), several ISCO values were recoded in EGP$$ as follows:
- ISCO 2470 becomes EGP=1.
- ISCO 2500 becomes EGP=2.
- ISCO 4300, 4400, 4500 become EGP=4.
- ISCO 7900 becomes EGP=7.
- ISCO 9910-9990 become EGP=9.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Please also see occupational status (ISEI$$) and occupational prestige scores (SIOPS$$, MPS$$).

Detailed description:

Ganzeboom, H. B. G. /Treiman, D. J. (1996) Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations In Social Science Research 25 201-239

Ganzeboom, H. B. G. /Treiman, D. J. (2003) Three Internationally Standardised Measures for Comparative Research on Occupational Status. In Hoffmeyer-Zlotnik, J. H. P. Wolf, C. (eds.) Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York Kluwer Academic/ Plenum Publishers. pp. 159–193.

SIOPS$$

Variable label	**"Treimans Standard Int. Occupation Prestige Score"**
Variable format	2-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable gives the occupational prestige score index for all employed persons. SIOPS$$ is based on ISCO-88 and was developed by Donald Treiman et al. The scale ranges from 6 to 78.

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Please also see occupational prestige scores (MPS$$), occupational status (ISEI$$), and occupational class (EGP$$).

Detailed description

Ganzeboom, Harry B.G. and Donald Treiman (1996) Internationally comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. In Social Science Research, Vol. 25, 201-239

MPS$$

Variable label	**"Magnitude-Prestige Scala – Wegener"**
Variable format	5-digit real
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable gives the occupational prestige score developed by Wegener (1988) for all employed persons. Like the SIOPS$$ prestige sore, Wegener's prestige scala measures a person's occupational prestige and was developed especially for use in the Federal Republic of Germany. MPS$$ is assigned based on the German Federal Statistical Office's occupational classification of 1992 (KLAS$$). The procedure has been documented in Frietsch and Wirth (2001).

This question is asked – similar to the SOEP – on a biannual basis for those who did not change their job. Hence, this information is written forward in following waves.

Please also see occupational prestige scores (SIOPS$$), occupational status (ISEI$$), and occupational class (EGP$$).

Detailed description:

Wegener, Bernd (1988) Kritik des Prestiges, Opladen.

Frietsch, Rainer, and Heike Wirth (2001) Die Übertragung der Magnitude-Prestigeskala von Wegener auf die Klassifikation der Berufe. In ZUMA Nachrichten 48 (Jg.25) 139-165.

AUTONO$$

Variable label	**"Autonomy in occupational activity"**
Value label	AUTONO$$ (0) apprentice, intern, unpaid trainee (1) low autonomy (2) low-medium autonomy (3) medium autonomy (4) medium-high autonomy (5) high autonomy
Variable format	1 digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable gives the occupational autonomy for all employed persons. It offers an alternative to the ISCO-based scales on occupational status (ISEI$$), class (EGP$$), or prestige (SIOPS$$). AUTONO$$ is the simplest variable based on the scales of "occupational position" in terms of its construction, and strongly correlated with the Treiman Prestige Scale (SIOPS$$).

The basis for the "autonomy in occupational activity" scale is the classification of occupational position. Self-employed persons are categorized according to the size of the company (with the exception of farmers, who are all classified within the same category of autonomy, independent of farm size in hectares). Civil servants are differentiated according to the civil service laws defining each kind of activity and the amount of autonomy connected to it. Workers are differentiated according to their vocational training, and thus categorized hierarchically according to the different tasks they can be expected to carry out and the different amounts of responsibility associated with each task. Similarly, salaried employees are classified according to how differentiated their tasks are and how much responsibility is associated with each.

The value "1" is assigned mainly to manual workers with a low level of status and a low level of autonomy. Group 2 encompasses work in production, services demanding a minimal level of specialization, and farm work. Activities that require completion of the middle track of secondary education and entail a limited amount of responsibility are classified in Group 3. Group 4 includes activities carried out either with or without supervision that require a degree from a college of applied sciences or university, but are not very high in prestige. Managers and freelance academics are both placed in Group 5 (highest autonomy). Depending on the number of employees, self-employed are categorized in Group 3, Group 4, or Group 5.

Detailed description: Hoffmeyer-Zlotnik, Jürgen H.P., and Alfons J. Geis (2003) Berufs-klassifikation und Messung des beruflichen Status/ Prestige. In ZUMA-Nachrichten 52, Jg. 27, Mai 2003. pp. 125-138.

AGRHRS$$

Variable label	**"Agreed-upon weekly working hours"**
Variable format	3-digit real
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable is designed to offer annual data on agreed weekly working hours. The variable takes into account only those persons who were in dependent employment (not self-employed) at the time of the survey.
The value ("-2 Does not apply") is assigned to employees without set hours and to self-employed people, including self-employed farmers, freelancers, persons helping out in family businesses, and other self-employed persons.

For implausible answers (agreed weekly working time of more than 80 hours per week) we assign the value ("-3 Answer implausible").
The value is rounded off and gives the number of working hours as a decimal number.

Please also see ACTHRS$$ and OVRHRS$$ for other variables on working hours.

ACTHRS$$

Variable label	**"Actual weekly working hours"**
Variable format	3-digit real
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable is designed to offer annual data on actual weekly working hours (including overtime) for all persons employed at the time of the survey (including the self-employed). The data are obtained by asking respondents how many hours they work *on average* per week.

For implausible answers (actual weekly working hours of more than 80 hours per week), we assign the value ("-3 Answer implausible"). The variable is rounded off and gives the number of working hours as a decimal number.

Please also see AGRHRS$$ and OVRHRS$$.

OVRHRS$$

Variable label	**"Overtime per week"**
Variable format	3-digit real
$$ - Survey Years	$$=10, 11, 12, 13

Comment

This variable is designed to offer annual data on overtime per week for all persons in dependent employment at the time of the survey. The data are obtained by asking respondents how many overtime hours they worked *in the month before the survey*. The number of monthly overtime hours is then converted into weekly overtime by dividing the

number given by 4.3. Since OVRHRS$$ refers to weekly overtime during the *last month*, the number may deviate from the difference between *average* actual weekly working hours and the agreed weekly working hours.

The value ("-2 Does not apply") is assigned for self-employment, including self-employed farmers, freelancers, persons helping out in family businesses, and other self-employed persons.

For implausible answers (agreed-upon weekly working time or actual weekly working time of more than 80 hours per week in addition to more than 10 overtime hours per week) we assign the value ("-3 Answer implausible"). The value is rounded off and gives the number of overtime hours as a decimal number.

Please also see AGRHRS$$ and ACTHRS$$.

PARTP$$

Variable label	**"Partner indicator (in HH)"**
Value label	PARTP$$ (0) no partner (in HH), clearly (1) spouse, clearly (2) partner, clearly (3) probably spouse (4) probably partner (5) spouse or partner, probably (9) partner exists, identity unknown
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

Partner indicators have the purpose of clearly defining spouse- (married) and partner- (unmarried) relationships in FiD households and thus enabling analyses on the *couple* level. The variable PARTP$$ generated in this context reveals whether a person in a FiD household has a partner in that household, and if so, the type of relationship existing between the partners. Relationships with persons outside the FiD household are not covered by this variable.
Code 0 is automatically assigned to all persons born before 1993 or persons living in households in which there is clearly no partnership. These include (a) one-(adult)person households, (b) single-parent households, (c) household head living together with only one parent (or parent-in-law) and (d) youth (turning 17 in survey year).
Codes 1 to 5 define the actual relationships. To assign Codes 1 and 2, the partnership has to be clearly defined from the perspective of both partners. This implies agreement between both partners in pointing to the respective partner within the household. No other contradiction may occur, e.g. a different indication by variable $STELL (= relationship to head of household in ***$pbrutto***) or marital status in that wave. Those variables were also used to confirm or revise the partner indicator, so that the indicator turns to 3 or 4 where minor contradictions are found

but could be solved. If not solved, code -3 ("implausible answer") is assigned. Code 3, 4 or 5 are assigned if partners are associated only via variable $STELL (= relationship to head of household in ***$pbrutto***), e.g. the combination 0 (=head of household) and 1 (=spouse of household head) and amended by information on family status if available. Inconsistencies can occur between the answers provided by the two persons or between data on marital status and relationship to head of household. In those cases each person is examined individually within his or her household context and the marital history is taken into account. If the uncertainty remains, the codes 3 or 4 are assigned. Code 9 is assigned if at least two other household members may potentially be a particular person's partner and thus no clear determination of partnership can be made.
Note that the variable PARTP$$ corresponds to PARTZ$$ in SOEP.

PARTNO$$

Variable label **"Person ID number of partner"**
Variable format 8-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment Partner indicators have the purpose of clearly defining spouse- (married) and partner- (unmarried) relationships in FiD households and thus to make analyses on the couple level possible.
If PARTP$$ is coded 0 or 9, this person has no partner or the partner cannot be identified as such. The variable PARTNO$$ is assigned the missing code "-2" (does not apply) for these persons. An exception is made if a match of a couple can be clearly accomplished but both persons indicate to not be together any more with the matched partner, i.e. they live together but are separated.
If PARTP$$ is coded 1, 2, 3, 4 or 5, a partnership was defined and PARTNO$$ is assigned the value of the unchanging person ID number (=PERSNR) of the partner.
For analyses of partner relationships, this information can be used to clearly link all persons with their respective partners, and all information on both partners can also be stored in a common dataset. To give as much information as available, it is possible that there are partners or spouses identified, who are actually separated, but still live in the same household. In these cases, the PARTNO$$ is set, but the PARTP$$ is left at 0 (no partner).

COUPST$$

Variable label **"Partner status"**
Value label COUPST$$

(1) Married, spouse in household
(2) Married, spouse not in household
(3) Coupled, partner in household
(4) Coupled, partner not in household
(5) Single
(6) Registered same-sex partnership, living together

(7) Registered same-sex partnership, living separately

Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment Partner status describes the existence of a partner within or outside the household.
COUPST$$ is based on information given by the respective person on his or her current relationship. For those whose partner was identified within the household, partner status is counter-checked with the information given by the partner, relation to the head of household as well as future reports on a given relationship. Thus note, that partner status in wave 1 can be different between data distributions due to consistency checks using up-to-date information from wave 2. When contradictions are not solvable, code "-3" (implausible answer) is assigned.

COUPID$$

Variable label "**Couple identifier within household"**
Variable format 4-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment Partner indicators have the purpose of clearly defining spouse- (married) and partner- (unmarried) relationships in FiD households and thus to make possible analyses on the couple level. COUPID assigns a unique, unchangeable ID to an identified couple. For analyses of partner relationships, this information can be used to clearly link all persons with their respective partners, and all information on both partners can also be stored in a common dataset. COUPID$$ is completely consistent with the corresponding ***biomarsy*** and ***biocouply*** data. Note that the assigned COUPIDs do **not** follow a consistent rank order in time within a person's life course.

MARRST$$

Variable label **"Marital status"**
Value label MARRST$$
(1) Married
(2) Married, separated
(3) Single (never widowed/divorced)
(4) Divorced (most recent event)
(5) Widowed (most recent event)
(6) Spouse abroad
(7) Registered same-sex partnership
(8) Registered same-sex partnership, separated
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment Marital status is solely describing the institutional status of marriage at the time of the person interview. To check for the existence of a current relationship refer to COUPST$$.

Marital status is based on information given by the respective person on his or her current relationship as well as on retrospective information about previous relationships asked in the biography questionnaire. Information on marital status when a child was born (provided in the biography information) is not used here, so contradicting information to ***biobirth*** might still be possible. For those whose partner was identified within the household, marital status is counter-checked with the information given by the partner. Where contradictions can be found, indication of the person information is compiled if reasonable. If no information is available, the indication by position related to head of household is deferred. Remaining contradictions are solved using information on marriage status when a child was born as well as future reports on a given relationship. Note that marital status in wave 1 can be different between data distributions due to consistency checks using up-to-date information from wave 2. Marital status is only available for people who are interviewed in the respective wave.

NATION$$

Variable label	**"Citizenship – nationality"**
Value label	NATION$$ (see below)
Variable format	3-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

(1) Germany
(2) Turkey
(3) Ex-Yugoslavia
(4) Greece
(5) Italy
(6) Spain
(7) Ex-GDR (Only Country Of Origin)
(10) Austria
(11) France
(12) Benelux
(13) Denmark
(14) Great Britain
(15) Sweden
(16) Norway
(17) Finland
(18) USA
(19) Switzerland
(20) Chile
(21) Romania
(22) Poland
(23) Korea
(24) Iran
(25) Indonesia
(26) Hungary
(27) Bolivia
(28) Portugal
(29) Bulgaria
(30) Syria
(31) Czech Republic
(32) Russia
(33) Empty (was Kurdistan)
(34) Mexico
(35) Argentina
(36) Cap Verde Is.
(37) Benin
(38) Philippines
(39) Israel
(40) Japan
(41) Australia
(42) India
(43) Afghanistan
(44) Thailand
(45) Jamaica
(46) Saudi Arabia
(47) Ethiopia
(48) Columbia
(49) Ghana
(50) Bangladesh
(51) Venezuela
(52) Tunisia
(53) Mauritius
(54) Nigeria
(55) Canada
(56) New Zealand
(57) Tanzania
(58) Cyprus
(59) Cuba
(60) Iraq
(61) Brazil
(62) Monaco
(63) Hong Kong
(64) Peru
(65) Sri Lanka
(66) Nepal
(67) Morocco
(68) China
(69) Liechtenstein
(70) Iceland
(71) Ireland
(72) St. Lucia
(73) Moldavia
(74) Kazakhstan
(75) Albania
(76) Lebanon
(77) Kyrgyzstan
(78) Ukraine
(79) Algeria
(80) Mozambique
(81) Egypt
(81) Egypt
(82) Tajikistan
(83) Vietnam
(84) Somalia
(85) Pakistan
(86) South Africa
(87) UAE
(88) El Salvador
(89) Eritrea
(90) Jordan
(91) Turkmenistan
(92) Costa Rica
(93) Singapore
(94) Burkina Faso
(95) Zambia
(96) Ecuador
(97) Uzbekistan
(98) No Nationality
(99) Puerto Rico
(100) Laos
(101) Estonia
(102) Angola
(103) Latvia
(104) Malaysia
(105) Namibia
(106) Montenegro
(107) Belize
(108) Dominican Republic
(109) Nicaragua

(110) Kenya
(111) Libya
(112) Malta
(113) Botswana
(114) Haiti
(115) Trinidad-Tobago
(116) Luxembourg
(117) Belgium
(118) Holland
(119) Croatia
(120) Bosnia-Herzegovina
(121) Macedonia
(122) Slovenia
(123) Slovakia
(124) Paraguay
(125) Guinea
(126) Kuwait
(127) Ivory Coast
(128) Malaysia
(129) Samoa
(130) Azerbaijan
(131) Seychelles
(132) Belarus
(133) Uruguay
(134) Bahamas
(135) Uganda
(136) Oman
(137) Micronesia
(138) Mali
(139) Cameroon
(140) Kosovo-Albania
(141) Georgia
(142) Sudan
(143) Congo
(144) Togo
(145) Mongolia
(146) Lithuania
(147) Chad
(148) Armenia
(149) Kurdistan
(150) Liberia
(151) Yemen
(152) Palestine
(153) Freistaat Danzig
(154) Taiwan
(155) Turkmenistan
(156) Africa
(157) Guatemala
(158) Sierra Leone
(159) Panama
(160) East Timor
(161) Bahrain
(162) Senegal
163) Maldives
(164) Kabarday
(165) Serbia
(166) Gambia
(167) Honduras
(168) Montenegro
(169) Cambodia
(170) Surinam
(171) Guyana
(172) Caucasus
(173) Zimbabwe
(174) Madagascar
(175) Grenada
(222) Eastern Europe
(333) Other, not specified foreign country

Educational Variables in FiD

As in the SOEP, there are three categories of educational variables in the FiD distribution of $PGEN: those dealing with schooling degrees, those with vocational and university degrees as well as the recoded values for the time spent in education.

In general, the information on all education variables for FiD is taken from two sources 1) The biographical part of the questionnaire in question $L23 to $L40 and 2) the information asked about current education surveyed in every wave (as the question numbers are subject to change every year, we do not quote any specific questions here).

For schooling degrees (SCEDU$$, SCEDUE$$ and SCEDUA$$), we combine the questions on schooling attainment to identify the respondent's secondary or tertiary school degree. While the respondents from the youth questionnaire are included (age 16 and older), we do not cover elementary schooling degrees. Whether a degree was obtained in Germany (pre or post unification), the former GDR (SCEDUE$$) or in another country (SCEDUA$$), is determined through the questionnaire as well. Note that the abroad category is only possible if the person has already returned to Germany to participate in FiD.

Higher educational degrees are covered in VCDEG$$, VCDEGE$$, VCDEGA$$, VCNONE$$ and COLLEG$$. Similar to the schooling attainment, they are based on the record in the $LELA files and updated by the more recent information if necessary. The information on attainment in the GDR (VCDEGE$$) and abroad (VCDEGA$$) comes from the biographical part, and – for the abroad category – also applies only, if the respondent has already returned to Germany.

Note that educational degrees coded from previous waves are written forward, unless the current wave contains new information. Previous waves are only updated if the new information also applies to that wave. E.g., if a respondent finishes a degree in wave 2, wave 1 data will remain on the old value. If it turns out, that the degree was already obtained before the wave 1 interview, wave 1 information is updated.

SCEDU$$

Variable label	**"Secondary/tertiary school degree"**
Value label	SCEDU$$
	(1) Basic-track sec. school (9th grade)
	(2) Interm.-track sec. school (10th grade)
	(3) Technical secondary school (12th grade)
	(4) Academic-track sec. school (13th grade)
	(5) Other graduation diploma
	(6) Left school without graduating
	(7) Not yet graduated
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment All respondents are asked about diplomas/degrees attained for completion of secondary/tertiary education in the biographical section of the questionnaire. While the variables SCEDUE$$ and SCEDUA$$ provide school degrees for respondents educated in the former GDR or in a foreign country, respectively, this variable combines all school education information. This data will be regularly updated to take into account any changes in highest diploma/degree attained.

SCEDUE$$

Variable label	**"Secondary school degree/diploma East Germany"**
Value label	SCEDUE$$
	(1) completion of 8th grade
	(2) completion of 10th grade
	(3) college entrance exam
	(4) other degree/diploma
	(5) dropout, no degree/diploma
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment As a supplement to the variable SCEDU$$ the highest secondary school degree/diploma in East Germany is provided as a separate variable. This information only originates from the biographical part of the questionnaire.

SCEDUA$$

Variable label	**"Secondary school degrees/diplomas abroad"**
Value label	SCEDUA$$
	(1) secondary school, no degree/diploma attained
	(2) secondary school, degree/diploma attained
	(3) vocational school
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment — As a supplement to the variable SCEDU$$, this variable provides annually updated data on the highest secondary school degree/diploma attained abroad.

VCDEG$$

Variable label — **"Vocational degree attained"**
Value label — VCDEG$$
(1) apprenticeship
(2) vocational school
(3) [health care school]
(4) technical school
(5) civil service training
(6) other training
Variable format — 1-digit integer
$$ - Survey Years — $$=10, 11, 12, 13

Comment — All respondents in all subsamples are asked about vocational degrees attained the first time they participate in FiD. This data is updated annually. While VCDEGE$$ and VCDEGA$$ provide information on vocational degrees obtained in the former GDR and abroad, respectively, VCDEG$$ provides this information on all respondents. VECDEG$$ captures the highest vocational degree attained, and in case of multiple degrees, VCDEGE$$ and VCDEGA$$ may not correspond with VCDEG$$.
Note that code (3) "health care school" is a category present in the SOEP, which was kept in FiD to allow direct comparisons. In later waves of the SOEP, this category is integrated in category (4).

VCDEGE$$

Variable label — **"Vocational degree attained – East"**
Value label — VCDEGE$$
(1) vocational training
(2) master craftsman
(3) engineering, technical degree
(4) other training
Variable format — 1-digit integer
$$ - Survey Years — $$=10, 11, 12, 13

Comment — To supplement the variable VCDEG$$ the highest secondary school degree/diploma in East Germany is provided as a separate variable.

VCDEGA$$

Variable label — **"Vocational degree abroad"**
Value label — VCDEGA$$
(1) on-the-job training
(2) vocational training
(3) vocational school
(4) college

(5) other
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment As a supplement to the variable VCDEG$$, this variable gives (and updates) the highest-level vocational degree attained abroad.

VCNONE$$

Variable label **"No vocational degree"**
Value label VCNONE$$
(1) no vocational degree
(2) still doing an apprenticeship
(3) still in university
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment In connection with the question about vocational degrees (VCDEG$$ and VCDEGE$$), all first-time respondents to all subsamples are explicitly asked, whether they (still) *do not* possess a vocational degree. In subsequent years, this data will be carried forward or updated. The variable has the Missing Value Code "-2 Does not apply", if one of the other two variables on vocational degree has a positive value.

COLLEG$$

Variable label **"Completed college education"**
Value label COLLEG$$
(1) technical college
(2) university, technical university
(3) college abroad
(4) engineering, technical school (East)
(5) university (East)
(6) doctorate degree
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment All respondents in all subsamples are asked about completed college education the first time they participate in FiD. To generate the variable, the different degrees/diplomas are integrated. Note that code "3 college abroad" is only assigned if the person reports to have obtained a degree when studying abroad. This is different to FiD version 1.2, where simply being abroad during college gave a code 3.

TIMEDU$$

Variable label **"Amount of education or training (in years)"**
Variable format 2-digit real
$$ - Survey Years $$=10, 11, 12, 13

Comment

The following statements describe the standard computation for schooling (including years of secondary vocational education).
The re-coding is not very complicated. For example, special schools for health care professions or other kinds of specialized schools are included in the "technical school" label. However, in Germany, this code is the one commonly used when earnings functions based on human capital theory are estimated.

Computation

The $BILZEIT variables are computed using the education variables provided by the $PGEN-files. The computation code is set to the following logic

no degree	7	years
lower school degree	9	years
intermediary school	10	years
degree for a professional college	12	years
high school degree	13	years
other	10	years

additional occupational training (includes universities)

apprenticeship	1.5	years
technical schools (incl. health)	2	years
civil servants apprenticeship	1.5	years
higher technical college	3	years
university degree	5	years

Detailed description:

Helberger, Christof (1988) Eine Überprüfung der Linearitätsannahme der Humankapitaltheorie. In H.-J. Bodenhöfer (ed.) Bildung, Beruf, Arbeitsmarkt, pp. 151-170, Berlin.
Schwarze, Johannes (1991) Ausbildung und Einkommen von Männern - Einkommensfunktionsschätzungen für die ehemalige DDR und die Bundesrepublik Deutschland. In Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, (24), pp. 63-69.

ISCED$$

Variable label **"ISCED-1997 classification"**
Value label ISCED$$
(0) in school
(1) inadequately
(2) general elementary
(3) middle vocational
(4) vocational + Abitur
(5) higher vocational
(6) higher education
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment

To make the educational degrees and diplomas attained in different countries comparable, for all respondents an educational variable

(ISCED$$) is generated retroactively from 1984 on using the international classification scheme ISCED-1997 (International Standard Classification of Education). It creates the highest degree/diploma attained, taking into account degrees and diplomas attained in both general schooling and in vocational and university education. Here the higher-level vocational and university override lower-level school diplomas. Persons who, for example, have no values for the variables on secondary school degrees/diplomas but state that they have a university degree are placed in the highest ISCED category.
Please note that, due to a lack of more detailed information on tertiary degrees -- in particular on promotion -- we include all tertiary degrees in our ISCED category 6. Thus, the ISCED variable provided here is not comparable one-to-one with the ISCED levels as defined by the OECD, since we have included the original ISCED level 5A in our ISCED category 6. See below for more details.

Computation

The ISCED$$ variables are computed using the education variables provided by the $PGEN-files. For this we use the variables on secondary degrees/diplomas (SCEDU$$) and secondary degrees/diplomas abroad (SCEDA$$), and the occupational education variables “vocational degree” (VCDEG$$), “university degree” (COLLEG$$) and “vocational degree abroad” (VCDEGA$$). We refrained from integrating the GDR-specific educational degrees/diplomas (SCEDUE$$ und VCDEGE$$).

ISCED FiD	SCEDU Schulabschluss	SCEDUA Schule im Ausland		VCDEG Berufliche Ausbildung	VCDEGA Berufsbildung im Ausland	COLLEG Hochschul-abschluss	ISCED97 – OECD
0	7.Noch kein Abschluss						0 noch kein Abschluss
1	5.Anderer Abschluss 6.Ohne Abschluss verlassen	1.Pflichtschule ohne Abschluss					1 ohne Abschluss verlassen
2	1.Hauptschul-abschluss 2.Realschulabschluss	2.Pflichtschule mit Abschluss					2 Haupt/Real-schulabschluss
3	3.Fachhochschulreife 4.Abitur	3.Weiterfuehrende Schule	|	1.Lehre 2.Berufsfachschule 6.Sonstiger Abschluss	2.Betriebliche Ausbildung 3.Berufsbildende Schule		3 Beruflicher/Real-schulabschluss oder (Fach)Abitur
4	3.Fachhochschulreife 4.Abitur	3.Weiterfuehrende Schule	**&**	1.Lehre 2.Berufsfachschule 6.Sonstiger Abschluss	2.Betriebliche Ausbildung 3.Berufsbildende Schule		4 Beruflicher Abschluss nach absolvierter allgemeinb. Schule
5	3.Fachhochschulreife 4.Abitur	3.Weiterfuehrende Schule	 **&**	3.Schule Gesundheitsw.(-99) 4.Fachschule, Meister 5.Beamtenausbildung			5B höherer beruflicher Abschluss
6					4.Hochschule	1.Fachhochschule 2.Universitaet, TH 3.Hochschule im Ausland 4.Ingenieur/Fach-schule (Ost) 5.Hochschule (Ost)	5A + 6 FH oder Universität + Promotion

Detailed description: OECD (1999) Classifying Educational Programmes Manual for ISCED-97 Implementation in OECD Countries. Paris 1999.

CASMIN$$

Variable label **"Highest degree/diploma according to CASMIN"**
Value label CASMIN$$

(0) (0) in school'
(1) (1a) inadequately completed '
(2) (1b) general elementary school'
(3) (1c) basic vocational qualification'
(4) (2a) intermediate general qualification'
(5) (2b) intermediate vocational'
(6) (2c_gen) general maturity certificate'
(7) (2c_voc) vocational maturity certificate'
(8) (3a) lower tertiary education'
(9) (3b) higher tertiary education'

Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment As an alternative to $ISCED, a second educational variable is generated ($CASMIN) that also enables comparison with international educational degrees/diplomas. Based on the modified CASMIN classification scheme (Comparative Analysis of Social Mobility in Industrial Nations), this variable has been computed retroactively from 1984 on for all respondents. Taken into account are both secondary-level and university/college-level degrees and diplomas. As with $ISCED, the higher-level occupational degrees override the lower-level secondary school degrees.

Computation The ISCED$$ variables are computed using the education variables provided by the $PGEN-files. For this we use the variables on secondary degrees/diplomas (SCEDU$$) and secondary degrees/diplomas abroad (SCEDA$$), and the occupational education variables "vocational degree" (VCDEG$$), "university degree" (COLLEG$$) and "vocational degree abroad" (VCDEGA$$). We refrained from integrating the GDR-specific educational degrees/diplomas (SCEDUE$$ und VCDEGE$$).

CASMIN	SCEDU Schulabschluss	SCEDUA Schule im Ausland		VCDEG Berufliche Ausbildung	VCDEGA Berufsbildung im Ausland	COLLEG Hochschulabschluss
0 – 0	7. Noch kein Abschluss					
1 – 1a	5. Anderer Abschluss 6. Ohne Abschluss verlassen	1. Pflichtschule ohne Abschluss				
2 – 1b	1. Hauptschulabschluss	2. Pflichtschule mit Abschluss				
3 – 1c	1. Hauptschulabschluss 5. Anderer Abschluss 6. Ohne Abschluss verlassen	2. Pflichtschule mit Abschluss	**&**	1-6. Beruflicher Bildungsabschluss	2. Betriebl. Ausbildung 3. Berufsbild. Schule	

4 – 2a	2. Realschulabschluss	3. Weiterfueh-rende Schule				
5 – 2b	2. Realschulabschluss	3. Weiterfueh-rende Schule	**&**	1-6. Beruflicher Bildungsabschluss	2. Betriebl. Ausbildung 3. Berufsbild. Schule	
6 – 2c_gen	3. Fachhochschulreife 4. Abitur					
7 – 2c_voc	3. Fachhochschulreife 4. Abitur		**&**	1-6. Beruflicher Bildungsabschluss	2. Betriebl. Ausbildung 3. Berufsbild. Schule	
8 – 3a						1. Fachhochschule
9 – 3b					4. Hochschule	2. Universitaet, TH 3. Hochschule im Ausland 4. Ingenieur/Fachsch. Ost 5. Hochschule (Ost)

Detailed description: The original version is described in König, W./Lüttinger, P./Müller, W. (1988) A Comparative Analysis of the Development and Structure of Educational Systems. Methodological Foundations and the Construction of a Comparative Educational Scale. CASMIN Working Paper No. 12. Mannheim Universität Mannheim.
For the modified version see Brauns, H./Steinmann, (1999) Educational Reform in France, West-Germany and the United Kingdom Updating the CASMIN Educational Classification. In ZUMA Nachrichten, Jg. 23, H. 44, pp. 7-44.

LABGRO$$

Variable label "**Current gross labor income in euros (generated)"**
Variable format 5-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment The variable LABGRO$$ represents the imputed current gross labor income generated for all FiD respondents, who are employed in the respective wave. LABGRO$$ is part of the individual imputation process conducted by multiple imputations in FiD (see documentation for details). Even though we recommend using all five implicates available in the dataset ***$mipinc***, we include the first implicate for convenience reasons in ***$pgen***.
The imputations for LABGRO$$ are based on the variable LABNET$$. The difference between the two is imputed, and in case of a missing value in *either* of the two original variables, LABGRO$$ is imputed by summing up the imputed values of LABNET$$ and the imputed difference between LABNET$$ and LABGRO$$. This procedure leads to slightly more imputed values, but guarantees that the gross income does not fall below the net income.

Imputed values are flagged (IMPGRO$$).

IMPGRO$$

Variable label	**"Imputation flag for LABGRO$$"**
Value label	IMPGRO$$ (0) observed value (1) imputed value
Variable format	1-digit Integer
$$ - Survey Year	$$=10, 11, 12, 13
Comment	The variable IMPGRO$$ identifies imputations of item-nonresponse in the variable LABGRO$$ (current gross labor income). Note that there are slightly more imputations here than there are missing values due to the imputation process in LABGRO$$.

LABNET$$

Variable label	**"Current net labor income (generated) in euros"**
Variable format	5-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	The variable LABNET$$ represents the imputed current gross labor income generated for all FiD respondents, who are employed in the respective wave. LABNET$$ is part of the individual imputation process conducted by multiple imputations in FiD (see documentation for details). Even though we recommend using all five implicates available in the dataset ***$mipinc***, we include the first implicate for convenience reasons in $PGEN. Imputed values are flagged (IMPNET$$).

IMPNET$$

Variable label	"**Imputation flag for LABNET$$"**
Value label	IMPNET$$ (0) observed value (1) imputed value
Variable format	1-digit Integer
$$ - Survey Year	$$=10, 11, 12, 13
Comment	The variable IMPNET$$ identifies imputations of item-nonresponse in the variable LABNET$$ (current net labor income).

FIELD$$

Variable label	**"Field of tertiary education"**
Value label	FIELD$$ 1-98 (see below)
Variable format	2-digit Integer
$$ - Survey Year	$$=10, 11, 12, 13

(1) Applied Linguistics and Cultural Studies
(2) Protestant Theology
(3) Catholic Theology
(4) Philosophy
(5) History
(6) Library Science, Archival Studies, Journalism
(7) Literary Studies, Linguistics
(8) Classical Philology, Modern Greek
(9) German Philology
(10) English Studies
(11) Romance Philology
(12) Slavonic Studies
(13) Non-European Languages and Cultural Studies
(14) Cultural Studies
(15) Psychology
(16) Educational Science
(17) Special Education
(22) Sport Science
(23) Law, Economics, Social Sciences
(24) Regional Studies
(25) Political Science
(26) Social Sciences
(27) Social Work
(28) Law
(29) Public Management and Governance
(30) Economics
(31) Business Administration and Engineering
(32) Mathematics, Natural Sciences
(36) Mathematics
(37) Computer Science
(38) Physics, Astronomy
(39) Chemistry
(40) Pharmacology
(41) Biology
(42) Geosciences
(43) Geography
(48) Health Sciences
(49) Medicine (without dentistry)
(50) Dentistry
(51) Veterinary Science
(57) Landscape Conservation, Environmental Design
(58) Agricultural Sciences, Food and Beverage Science
(59) Forest Management, Wood Management
(60) Nutrition and Home Economics
(61) Engineering
(62) Mining and Metallurgy
(63) Mechanical Engineering
(64) Electrical Engineering
(65) Traffic Engineering, Nautical Science
(66) Architecture, Interior Design
(67) Regional Planning
(68) Civil Engineering
(69) Surveying and Mapping
(74) Art, Aesthetics
(75) Fine Arts
(76) Design
(77) Performance, Film and Television, Theater
(78) Music, Musicology
(83) Outside the structure of the university system
(98) Not categorizable

Comment

The variable is designed to provide information on the field of education of tertiary degrees which adds details to the information recorded in the variable COLLEG$$. While the latter variable records if a person holds a degree FIELD$$ contains more detailed information on the type of the degree. The data of the generated variable FIELD$$ stem from two sources: 1. Person questionnaire: Each year respondents are asked if they have left education since the beginning of the year prior to the survey and which degrees they have obtained. This part of the questionnaire contains an open question on the type and the field of newly obtained tertiary degrees. This information is coded and used for the generation of the variables FIELD$$. 2. Biography questionnaire: Similar information is collected from respondents who fill in the biography questionnaire (usually during the first two years of participation in the panel). In contrast to the information from the person questionnaire the questions do not refer to currently obtained degrees but to degrees obtained during the time before being part of the FID sample.

In the variable FIELD$$ we combine these two types of information. Information on the data source is stored in the variable FDT_F$$.

Each year the variable contains the most recently collected information. Take for instance a person for whom we have observed a first degree in sociology in 2010 and a second degree in economics in 2012. For this person the variables FIELD$$ would be filled as follows:

2010-2011:	26	political/social science
2012-today	30	economics

If you want to take into account that a person holds two degrees you have to combine the information from all available years. However, only a minority of the population holds more than one tertiary degree. In very few cases we encounter the problem that a respondent provides information on two different degrees in one survey year. This only happens in years when respondents fill in the person as well as the biography questionnaire. In these cases we prioritize the information from the person questionnaire as it refers to the current situation while the biography questionnaire contains retrospective information. Furthermore, there are cases who report an applied university degree and a university degree in the biography questionnaire. In these cases, the variable contains information on the university degree only.

The variable is coded according to the classification on fields of education ("Fächergruppen") provided by the Statistisches Bundesamt (2009).

Detailed description Stat. Bundesamt (2009): Bildung und Kultur. Studierende an Hochschulen, Fachserie 11 Reihe 4.1, Wiesbaden: 446ff, Übersicht 1: „Fächergruppen, Studienbereiche und Studienfächer".

DEGREE$$

Variable label **"Type of tertiary degree"**
Value label DEGREE$$
11-98 (see below)

Variable format 2-digit Integer
$$ - Survey Year $$=10, 11, 12, 13

(11) Magister
(12) Diplom (University)
(13) Bachelor
(14) Master
(15) 1st State Examination
(16) Other state examination
(21) Diplom (at technical college, technical college for administration)
(22) Bachelor (at technical college, technical college for administration)
(31) Master (at technical college, technical college for administration)
(32) Teacher training,BA,MA at elementary, lower secondary schools/primary level
(33) Teacher training,BA,MA at 2ndary level 1/elementary schools/primary level
(34) Teacher training,BA,MA at intermediate scndry schools/scndry level I
(35) Teacher training, BA, MA at secondary level II and I
(36) Teacher training,BA,MA at academic 2ndry schools,2ndry levl 2,genrl school
(37) Teacher training, BA, MA at special needs schools

(38) Teacher training, BA, MA at vocational schools
(41) Teacher training, other
(42) Academic degree in the arts
(43) Doctorate
(44) Post-doctoral dissertation (Habilitation)
(45) Other degree
(98) Not categorizable

Comment: The variable is designed to provide information on the type of tertiary degree (e.g., Diploma, Bachelor, Master) which adds details to the information recorded in the variable COLLEG$$. While the latter variable records if a person holds a degree DEGREE$$ contains more detailed information on the type of the degree. The data of the generated variable DEGREE$$ stem from two sources: 1. Person questionnaire: Each year respondents are asked if they have left education since the beginning of the year prior to the survey and which degrees they have obtained. This part of the questionnaire contains an open question on the type and the field of newly obtained tertiary degrees. This information is coded and used for the generation of the variables DEGREE$$. 2. Biography questionnaire: Similar information is collected from respondents who fill in the biography questionnaire (usually during the first two years of participation in the panel). In contrast to the information from the person questionnaire the questions do not refer to currently obtained degrees but to degrees obtained during the time before being part of the FID sample.

In the variable DEGREE$$ we combine these two types of information. Information on the data source is stored in the variable FDT_F$$.

Each year the variable contains the most recently collected information. Take for instance a person for whom we have observed first an applied university diploma in 210 and a university diploma in 2012. For this person the variables DEGREE$$ would be filled as follows:

2010-2011:	21	diploma (applied university)
2012-today	12	diploma (university)

If you want to take into account that a person holds two degrees you have to combine the information from all available years. However, only a minority of the population holds more than one tertiary degree. In very few cases we encounter the problem that a respondent provides information on two different degrees in one survey year. This only happens in years when respondents fill in the person as well as the biography questionnaire. In these cases we prioritize the information from the person questionnaire as it refers to the current situation while the biography questionnaire contains retrospective information. Furthermore, there are cases who report an applied university degree and a university degree in the biography questionnaire. In these cases, the variables contain information on the university degree only.

The variable is coded according to a slightly collapsed version of the classification on types of tertiary degrees (“Prüfungsgruppen und Abschlussprüfungen”) provided by the Statistisches Bundesamt (2009).

Detailed description Stat. Bundesamt (2009): Bildung und Kultur. Studierende an Hochschulen, Fachserie 11 Reihe 4.1, Wiesbaden: 449ff, Übersicht 2: „Prüfungsgruppen und Abschlussprüfungen“.

TRAINA$$

Variable label **“Apprenticeship – two-digit occupation KldB92”**
Value label TRAINA$$
1-99 (see below)

Variable format 2-digit Integer
$$ - Survey Year $$=10, 11, 12, 13

(1) Agricultural Occupations (Crops)
(2) Agricultural Occupations (Livestock)
(3) Administrative/Advisory/Technical Specialist In Agriculture
(5) Horticultural Occupations
(6) Forestry and Hunting Occupations
(7) Mineworkers
(8) Mineral Exploitation and Processing
(10) Stonemasons
(11) Manufacturers of Construction Materials
(12) Ceramicists
(13) Glass Manufacturing Occupations
(14) Chemical Industry Occupations
(15) Plastics Manufacturing Occupations
(16) Paper Manufacturing and Processing
(17) Printing Occupations
(18) Wood and Woodworking, Wickerwork Occupations
(19) Occupations in Iron and Steelmaking, Semi-Finished Products
(20) Pouring and Casting Occupations
(21) Metal Processing Occupations (Chipless Forming)
(22) Metal Processing Occupations (Chip Forming)
(23) Occupations in the Metal Surface Treatment and Finishing Industry
(24) Metal Compounding Occupations
(25) Metal and Plant Construction Occupations
(26) Sheet Metal Manufacturing Occupations
(27) Mechanical Engineering and Maintenance Occupations
(28) Automotive and Aircraft Manufacturing and Maintenance Occupations
(29) Tool and Die Making Occupations
(30) Precision Engineering and Related Occupations
(31) Electrical Occupations
(32) Metalworkers, not otherwise mentioned
(33) Occupations in Spinning
(34) Occupations in Textile Production
(35) Occupations in Textile Processing
(36) Occupations in Textile Finishing
(37) Occupations in Leather Production and Processing
(39) Occupations in Baking, Confectionery, and Candy Production
(40) Butchers
(41) Chefs
(42) Occupations in Beverages, Alcohol, and Tobacco Manufacturing
(43) Other Occupations in Nutrition
(44) Occupations in Structural Engineering
(46) Occupations in Civil Engineering
(47) Unskilled Construction Workers
(48) Construction Finishing Occupations
(49) Interior Decorator, Upholsterers
(50) Occupations in Woodworking and Polymer Processing
(51) Painters, Varnishers, and Related Occupations
(52) Quality Control Inspectors, Mailing, and Dispatching Staff
(53) Unskilled Laborers, responsibilities not specified
(54) Machine Operators, not otherwise specified
(55) Machine Fitters, not otherwise mentioned
(60) Engineers, not otherwise mentioned
(61) Chemists, Physicists, Mathematicians
(62) Technicians, not otherwise mentioned
(63) Technical Specialists
(64) Technical Drafting, related occupations
(65) Industrial, Factory, Training Foremen/Forewomen
(66) Salespeople
(67) Wholesale and Retail Salespeople, Purchasing and Sales Staff
(68) Product Sales Staff, not otherwise specified, Sales Representatives
(69) Banking, Savings Association, Insurance Specialists
(70) Service Industry and Related Occupations
(71) Surface Transport Occupations
(72) Water and Air Traffic Occupations
(73) Communications Occupations
(74) Stock Clerks, Warehouse and Transport Workers
(75) Occupations in Management, Consulting, and Auditing

(76) Members of Parliament, Administrative Staff
(77) Invoicing Officers, Computer Scientists
(78) Office Occupations, Commercial Staff, not otherwise mentioned
(79) Service and Guard Occupations
(80) Security Occupations, not otherwise mentioned
(81) Occupations in Law and Law Enforcement
(82) Journalism, Translation, Library Science, and Similar OccupationsBerufe in der Unternehmensleitung, -beratung und -prüfung
(83) Artistic and Related Occupations
(84) Doctors, Pharmacists
(85) Other Health Care Occupations
(86) Social Occupations
(87) Teachers
(88) Occupations in the Humanities and Natural Sciences
(89) Pastoral Occupations
(90) Personal Care Occupations
(91) Occupations in Hotels and Hospitality
(92) Occupations in Domestic and Nutritional Science
(93) Cleaning and Waste Management Occupations
(96) Others
(97) Family members providing assistance,not in agriculture,not otherw. mntnd
(98) Workers, (still) without specific occupation
(99) Workers, responsibilities not specified

Comment: The variable is designed to provide information on the occupation of vocational training which adds details to the information recorded in the variable VCDEG$$. In addition to the variable TRAINA$$ we provide the variables TRAINB$$, TRAINC$$ and TRAIND$$. All these variables record the occupation of vocational training. The difference is that TRAINA$$ contains information on vocational training within the German dual system which combines firm-based and school-based training (apprenticeship). TRAINB$$ is designed to provide information on the occupation of full-time school based vocational training. TRAINC$$ contains information on level vocational training (e.g., Meister, Techniker). TRAIND$$ is designed to provide information on the occupation of civil servant training ("Beamtenausbildung"). We describe in brief detail the construction of the variable TRAINA$$. TRAINB$$, TRAINC$$ and TRAIND$$ are constructed in an analogous manner.

The data of the generated variable TRAINA$$ stem from two sources: 1. Person questionnaire: Each year respondents are asked if they have left education since the beginning of the year prior to the survey and which degrees they have obtained. This part of the questionnaire contains an open question on the type and the field of newly obtained tertiary degrees. This information is coded and used for the generation of the variables TRAINA$$. 2. Biography questionnaire: Similar information is collected from respondents who fill in the biography questionnaire (usually during the first two years of participation in the panel). In contrast to the information from the person questionnaire the questions do not refer to currently obtained vocational qualifications but to qualifications obtained during the time before being part of the FID sample.

In the variable TRAINA$$ we combine these two types of information. Information on the data source is stored in the variable FDT_F$$.

Each year the variable contains the most recently collected information. Take for instance a person for whom we have observed a first vocational qualification as an electrician in 2010 and a second

qualification as a car mechanic in 2012. For this person the variables TRAINA$$ would be filled as follows:

2010-2011:	31	electrical occupation
2012-today	28	automotive/flight industry occupation

If you want to take into account that a person holds two vocational qualifications you have to combine the information from all available years. In few cases we encounter the problem that a respondent provides information on two different apprenticeships in one survey year. This only happens once, namely in years when respondents fill in the person as well as the biography questionnaire. In these cases we prioritize the information from the person questionnaire as it refers to the current situation while the biography questionnaire contains retrospective information.

The variable is coded according to the classification of occupations at two-digit level ("Berufsgruppen") provided by the Statistisches Bundesamt (1992).

Detailed description Hartmann/Schütz (2002): Die Klassifikation der Berufe und der Wirtschaftzweige im Sozio-oekonomischen Panel – Neuvercodung der Daten 1984 – 2001. Infratest Sozialforschung, München.

TRAINB$$

Variable label	**"Vocational school – two-digit occupation KldB92"**
Value label	TRAINB$$ 1-99 (see TRAINA$$)
Variable format	2-digit integer
$$ - Survey Year	$$=10, 11, 12, 13

Comment The variable is designed to provide information on the occupation of full-time school based vocational training (e.g., Berufsfachschule, Schule des Gesundheitswesens, Handelsschule). See the description of variable TRAINA$$ for more details on the construction and the values of the variable.

TRAINC$$

Variable label	**"Higher vocational school – two-digit occupation KldB92"**
Value label	TRAINC$$ 1-99 (see TRAINA$$)
Variable format	2-digit integer
$$ - Survey Year	$$=10, 11, 12, 13

Comment The variable is designed to provide information on the occupation of higher level vocational training (e.g., Meister, Techniker). See the description of variable TRAINA$$ for more details on the construction and the values of the variable.

TRAIND$$

Variable label	**"Civil servant training – two-digit occupation KldB92"**
Value label	TRAIND$$ 1-99 (see TRAINA$$)
Variable format	2-digit Integer
$$ - Survey Year	$$=10, 11, 12, 13
Comment	The variable is designed to provide information on the occupation of civil servant training ("Beamtenausbildung"). See the description of variable TRAINA$$ for more details on the construction and the values of the variable.

FDT_F$$

Variable label	**"Data source FIELD, DEGREE, TRAIN"**
Value label	FDT_F$$ (1) person questionnaire (2) person questionnaire (temporary drop-out) (3) biography questionnaire (4) various sources
Variable format	1-digit integer
$$ - Survey Year	$$=10, 11, 12, 13
Comment	This is a flag variable which provides information on the data sources used for the construction of the variables FIELD$$, DEGREE$$, TRAINA$$, TRAINB$$, TRAINC$$ and TRAIND$$ (see the description of the respective variables for details).

Documentation *$hgen*

Household-related status variables and generated variables

Mathis Fräßdorf (geb. Schröder) and Malisa Zobel

*This documentation is based on the comparable SOEP documentation on **$hgen** and has benefited from previous work of Joachim Frick, Markus Grabka, Jan Goebel, Peter Krause, Olaf Groh-Samberg and Christian Schmitt. Please understand that for readability reasons, we do not specifically cite and specify text that has been used directly from the SOEP document.*

General information

The ***$hgen*** datasets provide household level information for each wave of the survey. In some sense, the variables included in ***$hgen*** allow the researcher to have an easier access to household information. As the variable names and formats are consistent over all years, a combination of these files allows an easy comparison of different years. In addition, some variables are completely generated or imputed, and hence provide information which is not included in the regular questionnaire. The following documentation lists all variables in ***$hgen*** and provides information on their generating process. Information provided looks as follows for all variables:

Variable Label	Provides the label of the variable as it is given in the dataset. Variables are given in CAPTIAL letters, even though they might appear in small letters in the dataset. This is simply for readability.
Value Labels	*LBLNME* In case *VARNME* is categorical, *LBLNME* specifies the labels for each category, and the value labels are listed here. Note that the standard missing value labels (-1: No answer; -2: Does not apply; -3: Not valid) are not listed, but apply to all variables in this dataset.
Variable format	Specifies the format for each variable, e.g. “1-digit integer” or “string”.
$$ - Survey Years	Specifies the years for which the variable is provided. This is provided for 2000+, such that “10” refers to 2010, etc.
Comment:	Provides more detailed information on the generating process, also on the population the variable is specified for, if necessary. Here, variables used, changes between waves, or any other anomalies are mentioned and their relevance explained.

The variables described in the following are in part status variables in this sense: information collected once will be carried forward to subsequent years if no address change has taken place since the previous year. This is the case for: CNSTYR$$, CONDIT$$, SIZE$$, ROOM$$, EQPKIT$$, EQPSHW$$, EQPIWC$$, EQPHEA$$, EQPTER$$, EQPBAS$$, EQPGAR$$, EQPWAT$$, EQPTEL$$, EQPALM$$, EQPSOL$$, EQPAIR$$, MOVEYR$$, RSUBS$$ and REDUC$$.

Comparison with SOEP

The following variables, which are part of the SOEP distribution, are not part of FiD, as this information has not been collected. Whether the information will be included in a future wave of FiD, is not clear at the moment.

ACQUIS$$	Means of acquiring dwelling
EQPTEL$$	Dwelling has telephone
EQPIWC$$	Dwelling has indoor toilet
SUBSID$$	Government-subsidized housing payments

On the other hand, there are two variables for which information was collected for the first time (it was also collected in SOEP 2010):

EQPFHEA$$	Dwelling has floor heat
ELECTR$$	Amount of monthly electricity costs in Euro

If you have questions regarding ***$hgen*** data for the FiD-distribution, please contact Mathis Schröder.

List of Variables:

Contents

CNSTYR$$

Variable label	**"Year house was constructed"**
Value labels	CNSTYR$$ (1) before 1919 (2) 1919 to 1948 (3) 1949 to 1971 (4) 1972 to 1980 (5) 1981 to 1990 (6) 1991 to 2000 (7) 2001 or later (8) 2011 or later
Variable format	1 digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Classified statement of year the building was constructed in which a household lives at the time the survey was built.

CONDIT$$

Variable label	**"Condition of building"**
Value labels	CONDIT$$ (1) in good condition (2) partial renovation needed (3) major renovation needed (4) ready for demolition
Variable format	1 digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Respondent's assessment of the condition of the building.

SIZE$$

Variable label	**"Size of housing unit in square meters"**
Variable format	3-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Size of housing unit could be missing due to non-response. However, the value is imputed for all missing values using the household multiple imputation procedure (see imputation documentation for details). The first implicate of these imputations is used to impute SIZE$$. See FSIZE$$ for the imputation indicator. If, however, the household reported the size in any wave and never moved in between, the reported value is used instead of the imputation. This is not indicated by the imputation flag FSIZE$$.
See documentation on imputations for more detail on the multiple imputations.

FSIZE$$

Variable label **"Imputation flag for size of housing unit"**

Value labels	FSIZE$$ (0) observed value (1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

ROOM$$

Variable label	**"Number of rooms larger than 6 square meters"**
Variable format	2-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Number of rooms could be missing due to non-response. However, the value is imputed for all missing values using the household multiple imputation procedure (see imputation documentation for details). The first implicate of these imputations is used to impute ROOM$$. See FROOM$$ for the imputation indicator. If, however, the household reported the number of rooms in any wave and never moved in between, the reported value is used instead of the imputation. This is not indicated by the imputation flag FROOM$$.
See documentation on imputations for more detail on the multiple imputations.

FROOM$$

Variable label	**"Imputation flag for size of housing unit"**
Value labels	FROOM$$ (0) observed value (1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

SEVAL$$

Variable label	**"Adequacy of living space in the housing unit"**
Value label	SEVAL$$ (1) much too small (2) a bit too small (3) just right (4) a bit too large (5) much too large
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Assessment of adequacy by respondent (household head).

EQPKIT$$

Variable label	**"Dwelling has kitchen"**
Value label	EQPKIT$$ (1) yes

(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPSHW$$

Variable label **"Dwelling has indoor bath/shower"**
Value label EQPSHW$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPHEA$$

Variable label **"Dwelling has central heating"**
Value label EQPHEA$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPFHEA$$

Variable label **"Dwelling has floor heating"**
Value label EQPHEA$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment: This information has not been asked in the SOEP before 2010.

EQPTER$$

Variable label **"Dwelling has balcony/terrace"**
Value label EQPTER$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPBAS$$

Variable label **"Dwelling has basement"**
Value label EQPBAS$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPGAR$$

Variable label	**"Dwelling has garden"**
Value label	EQPGAR$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

EQPWAT$$

Variable label	**"Dwelling has hot water/boiler"**
Value label	EQPWAT$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

EQPALM$$

Variable label	**"Dwelling has alarm device"**
Value label	EQPALM$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

EQPSOL$$

Variable label	**"Dwelling has solar collector"**
Value label	EQPSOL$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

EQPAIR$$

Variable label	**"Dwelling has air conditioning"**
Value label	EQPAIR$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

EQPNRJ$$

Variable label	**"Dwelling has other alternative energy source"**
Value label	EQPNRJ$$
	(1) yes
	(2) no

Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

EQPLIF$$

Variable label **"Dwelling has lift"**
Value label EQPLIF$$
(1) yes
(2) no
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

MOVEYR$$

Variable label **"Year moved into dwelling"**
Variable format 4-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment: The information given in MOVEYR$$ was collected on an individual level in the biographical section of the person questionnaire. It was then collapsed on the household level, such that the individual living in the household the longest provided the year of moving in.
Beginning with survey year 2011 for households at their old address, data are carried forward. For new households and for old households that have moved, the variable is based on newly collected data. In case the information is missing and an old household has moved that year or the previous year, MOVEYR$$ will take the value of the year of the respective wave. The carrying forward of data entails the possibility that the year of moving into the new dwelling may lie before the year of birth of the oldest household member.

OWNER$$

Variable label **"Tenant or owner of dwelling"**
Value label OWNER$$
(1) Owner
(2) Main tenant
(3) Subtenant
(4) Tenant
(5) Institutional resident (Nursing home etc.)
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment: Code 4 is used if the original variable is coded as -1 ("No answer") but if at least one answer that is specific to tenants, was given. If the original variable is "-1" but at least one answer specific to owners was given, then the "-1" is recoded to "1" ("Owner"). Code 4 is also used if a change in ownership (from owner to tenant) has taken place, but no original information for OWNER$$ was given.

OSUBS$$

Variable label	**"Received government housing subsidies last year"**
Value label	OSUBS$$
	(1) yes
	(2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: This information is given for owners-occupiers only. OSUBS$$ contains information on cash housing subsidies received from the government during the year prior to the interview. This information is based solely on the respondent's report, no other information is used. If missing, the information is **not** carried forward.

REVAL$$

Variable label	**"Evaluation of rent paid"**
Value label	REVAL$$
	(1) very inexpensive
	(2) inexpensive
	(3) reasonable
	(4) slightly too expensive
	(5) too expensive
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: This information is given for tenant-occupiers only. It contains the subjective assessment by the respondent (household head). The corresponding information from the previous year is **not** carried forward longitudinally due to the possibility of changes in rent and income, residential moves, and change in the person responding.

RSUBS$$

Variable label	**"Government subsidized housing"**
Value label	RSUBS$$
	(1) yes, with subsidy
	(2) yes, with expired subsidy
	(3) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: This information is given for tenant-occupiers only. Beginning with survey year 2011 information will be carried forward from previous years for immobile households. The code "2" was introduced to indicate expired subsidization for subsequent waves.

REDUC$$

Variable label **"Rent-reduced dwelling"**

Value label	REDUC$$ (1) yes (2) no
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: This information is given for tenant-occupiers only. Beginning with survey year 2011, information will be carried forward from the previous years for old households residing at their old address; for new households and for old households that have moved, newly collected data will be used.

RENT$$

Variable label	**"Amount of rent, no heating, with utilities (EURO)"**
Variable format	4-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

HEAT$$

Variable label	**"Amount of monthly heating costs (EURO)"**
Variable format	4-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

UTIL$$

Variable label	**"Amount of monthly utility costs (EURO)"**
Variable format	4-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

ELECTR$$

Variable label	**"Amount of monthly electricity costs (EURO)"**
Variable format	4-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment for RENT$$, HEAT$$, UTIL$$, ELECTR$$
All four variables are only provided for tenants-occupants who pay rent (NORENT$$ not equal "1"). In the questionnaire, the information of all four variables is asked separately for all tenants. Respondents report their monthly payments for rent, heating, electricity and utilities. For heating the respondents further specify whether it is included in rent or not. RENT$$ is then generated depending on the answers to the inclusion question. If included, then the amount of heating costs is subtracted from the answer given in to the rent question.

If any of the above values is missing, they are part of the household multiple imputation procedure (see imputation documentation for details). Although we recommend using the multiple imputations to fully use their distributional properties, as elsewhere in ***$hgen***, we use the first of five implicates for those users who do not wish to deal with multiple imputations. All calculations regarding RENT$$ are done with the imputed first implicate as well. The variables FRENT$$, FHEAT$$, FUTIL$$, FELECTR$$ contain the respective imputation flags.

FRENT$$

Variable label	**"Imputation flag gross rent"**
Value labels	FSIZE$$
	(0) observed value
	(1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: This flag captures whether any component of the rent calculation was imputed. As RENT$$ is a combination of the rental value, heating costs and possibly the utility costs, the number of values including any imputations is larger than the number of imputed rents. The variable i_$hrent in the file ***$mihinc*** depicts how many of the actual rental values have been imputed.

FHEAT$$

Variable label	**"Imputation flag for heating costs"**
Value labels	FHEAT$$
	(0) observed value
	(1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

FUTIL$$

Variable label	**"Imputation flag for utility costs"**
Value labels	FUTIL$$
	(0) observed value
	(1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

FELECTR$$

Variable label	**"Imputation flag for electricity costs"**
Value labels	FELECTR$$
	(0) observed value
	(1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

NORENT$$

Variable label	**"Pays no rent"**
Value label	NORENT$$
	(1)Does not pay rent
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment: Coded as "1" if the household does not pay rent; e.g., if living space is provided by relatives at no cost. Note that in these cases, the information on gross cold rent (RENT$$), heating costs (HEAT$$), utilities (UTIL$$) and electricity (ELECTR$$) is coded "-2" ("not applicable").
Information will not be carried forward.

TYP1HH$$

Variable label **"Household typology (1-digit)"**
Value label TYP1HH$$
(1) 1-person HH
(2) Childless couple
(3) Single parent
(4) Couple with children <= 16 yrs.
(5) Couple with children > 16 yrs.
(6) Couple with children <= 16 and > 16 yrs
(7) Multiple generations HH
(8) Other combination
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

TYP2HH$$

Variable label **"Household typology (2-digit)"**
Value label TYP2HH$$
(11) 1-P-HH Man < 35
(12) 1-P-HH Man 35-<60
(13) 1-P-HH Man >=60
(14) 1-P-HH Woman < 35
(15) 1-P-HH Woman 35-<60
(16) 1-P-HH Woman >=60
(21) Childless couple
(31) Single parent + 1 child
(32) Single parent + 2 or more children
(33) Single parent + 1 EK.
(34) Single parent + 2 or more EK.
(35) Single parent + 2 (E)K.
(36) Single parent + 3 or more(E)K.
(41) Couple + 1 child
(42) Couple + 2 children
(43) Couple + 3 or more children
(51) Couple + 1 EK.
(52) Couple + 2 EK.
(53) Couple + 3 or more EK.
(61) Couple + 2 (E)K.
(62) Couple + 3 or more (E)K.
(71) 3-generation HH
(72) 4-generation HH
(73) Grandparents-Children HH
(81) Other combination without children

(82) Other combination + 1 or more children

Variable format 2-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment: Generated variable created by combining the relationships of all persons living in the household to the head of household (Variable $STELL in the file $PBRUTTO) at the time of the survey.
TYP1HH$$ is an aggregation of TYP2HH$$ (first column of the two-digit code). Single households are differentiated in TYP2HH$$ according to both gender and age.

Legend:

- K = children up to the age of 16;
- EK = adult children age 17 and older;
- (E)K = children both below and above age 16;
- 1-P-HH = one-person households.

HINC$$

Variable label **"Monthly net household income (EURO)"**
Variable format 5-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment: This variable contains the current monthly net household income asked for in the household questionnaire, always provided in Euros. Income is reported by the respondent (head of household).

AHINC$$

Variable label **"Adjusted monthly net household income (EURO)"**
Variable format 5-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment on AHINC$$
This variable is based on the current monthly net household income asked for directly in the household questionnaire ("screener"). Since everyone in FiD over the age of 16 is also interviewed personally, this income can be calculated based on the current individual monthly incomes of all household members. Possible underestimation in "screener" can thus be assessed and corrected. However, in the case of item-nonresponse in the original screener, the procedure is only used for households surveyed completely, without item-non-response on the variables in question.

For personal income, we use monthly net income (from dependent employment and self-employment), extra earnings, pensions, widow's pensions, unemployment benefits or relief, maintenance payments, early retirement payments, maternity benefits, BaFoeG (state higher education grants), military or civil service pay, compulsory child support, as well as other forms of support from the $P files.

Civil servants' pension income is taxed at the flat rate of 20% and multiple entries on the use of employment office services are corrected for by calculating a median value.

We add all the individual incomes of all interviewed household members, also adding to this sum all income from the household context (housing subsidies, child benefits, welfare and home nursing subsidies, social assistance, unemployment Benefit II or Social Benefit ("ALG II")).

When no answer was provided on net household income, we use the net household income calculated as described above, under the condition that all household members gave valid answers.

If the net household income generated in this way is higher than the household income stated in the questionnaire, we correct the value upwards. If the generated household income is lower, we stay with the value originally stated.

When no answer was given for the different components of income, we set the value of the particular component to zero. An overview of the different components (excluding net monthly income, which is available each year) is provided in the table below, in which "x" indicates that the particular component was taken into account in that particular year.

Year	Additional earnings	Old-age pensions	Widow's pensions	Unemployment benefits	Unemployment relief	Maintenance payments while in higher education	Maternity benefits	Bafög (state higher education grants)	Military or civil service pay	Compulsory child support	Support payments
2010	-	x	x	x	-	x	x	x	x	x	x
2011	-	x	-	x	-	x	x	x	x	x	x
2012	-	x	-	x	-	x	x	x	x	x	x
2013	-	x	-	x	-	x	x	x	x	x	x

I_HINC$$

Variable label **"Imputed monthly net household income (EURO)"**
Variable format 5-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment on I_HINC$$
FiD uses a household multiple imputation procedure which jointly imputes all necessary values for the household. See the documentation on multiple imputations for detailed information. The variables I1HINC$$ to I5HINC$$ are the result of these multiple imputations and are provided here in the wide format (i.e. the number of observations per household is not changed). Complete imputation results are provided in the dataset ***$mihinc***. The imputation flag FHINC$$ identifies households with imputed incomes.

Analyzing multiply imputed data
For analyzing multiple imputed data, you do not necessarily need special methods, however such tools exist and simplify the use of multiply imputed data. Below a short overview of some useful tools for various statistical packages is given. These tools estimate the parameters of a regression model by combining the estimates across the several replicates of imputation. Point estimates from multiple imputations are then the arithmetic mean of the several point estimates obtained from analysis on each imputed data. Standard errors are obtained by

combining the average of the squared standard errors of the several (m) estimates with the within- and between-imputation variance.

- Since Stata™ version 11.0, there is a complete *mi* module available with various useful tools for analyzing multiple imputed data, but also for imputation itself.
- Within SAS, PROC MIANALYZE combines the results of analyses on the data sets.
- IVEware is a set of routines that can be launched from SAS or run independently using data from many sources. You can use the IVEware module regress to perform multiple imputation analysis.

FHINC$$

Variable label	**" Imputation flag for household income"**
Value labels	FHINC$$ (0) observed value (1) imputed value
Variable format	1-digit Integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment:	FHINC$$ is a dummy variable indicating whether an observation was missing on HINC$$ and was therefore imputed or not.

NUTS$$

Variable label	**"NUTS-systematic 1 (Federal State)"**
Value label	NUTS$$ (1) Baden-Wuerttemberg (2) Bavaria (3) Berlin (4) Brandenburg (5) Bremen (6) Hamburg (7) Hesse (8) Mecklenburg-Western Pomera (9) Lower Saxony (10) North Rhine-Westphalia (11) Rhineland-Palatinate (12) Saarland (13) Saxony (14) Saxony-Anhalt (15) Schleswig-Holstein (16) Thuringia
Variable format	2-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	The NUTS$$ variable provides the basic classification of the German "Bundesländer".

Documentation *bioage01-10* Files

Generated variables from the Parent-Questionnaires

Alexander Raith, Nina Scherner, Malisa Zobel, Moritz Mannschreck, Linda Wittbrodt and Mathis Fräßdorf (geb. Schröder)

Table of Contents

List of Tables and Figures

Introduction

The ***bioage*** data files are generated using information collected in the "Parent-Questionnaires". There are six different "Parent-Questionnaires" (for an overview see table 3) and accordingly six different ***bioage*** data files[6] with the data file specific suffix indicating the age group of the children (e.g. children in ***bioage01*** are at most 23 months old, those in ***bioage02*** are between 24 and 35 months old, etc.).

The aim of the "Parent-Questionnaires" is to follow and observe future generations of the population in *Familien in Deutschland* (FiD) and collect all information in age specific files, even though the data come from different survey years. As we try to make this process as comprehensive and gap-free as possible, we start following and documenting the development of children in FiD households from birth onwards. Consequently, ***bioage*** data files contain information regarding child birth, pregnancy, child's health, child care and further aspects related to raising children (for an overview of subjects covered in ***bioage01-10***, see table 2). Since many questions overlap, all ***bioage*** data files are documented within this chapter, rather than dedicating each ***bioage*** data file a single documentation file. By doing so, we are able to provide an overview of all variables covered in the ***bioage*** data files.

The ***bioage*** files in FiD have their conceptual match in the ***bioage*** files known from the SOEP – however, the FiD parent questionnaires cover slightly more subjects than those in the SOEP (where they are named "MuKi"-files, for "Mother-Child" questionnaires). A direct comparison of FiD and SOEP content is provided in table 1. Also, FiD covers one age group not captured in the SOEP so far – the 1-2 year-olds. Starting in the data collection of 2012 (wave BC), the SOEP also collects information about 9-10 year olds. However, data about the children are only collected from one parent, such that less information is available per child.

The rules used to generate the variables from the questionnaires are consistent over the various ***bioage*** data files. For the most part, the variables in the ***bioage*** files are simply renamed from their respective parent-questionnaires. This ensures identity over the years, even if different questionnaires (in terms of question ordering or content) are used. The renaming also allows the important comparison across the different age groups, as variable names are identical (up to their prefix), if their contents match, see table 1. A few variables are generated by combining the information from two or more variables. The 'common origin' (i.e. the common question) of a range of variables is indicated by using the same stem

[6] You will find two ***bioage08*** and two ***bioage10*** data files, with suffix p1 and p2. We refer to those as one dataset here, respectively, as the variable content is identical. See below for the specifics of ***bioage08*** and ***bioage10***.

word (e.g. "MVMN"), followed by a number (e.g. "MVMN2"). In some cases those slight variations match the difference between questionnaires for younger and older children, in which case the letter "y" for younger (e.g. CHAR6Y).[7]

Additionally, this documentation provides detailed information on variables that were generated using information from other data files (such as the p-files). You can find an overview of the specific variables and the rules by which they were generated in section 6 of this chapter.

Note that while parents (see section 4 on specifics) are the respondents, data are organized according to the never-changing personal identification number of the child (PERSNR**)**. In case of ***bioage08*** and ***bioage10*** this leads to two possible observations for each child (one from the 'father respondent' and one from the 'mother respondent'), stored in two separate data files. As such, you will only find one observation per PERSNR in any ***bioage*** data file. The only exception to this rule is the dataset ***bioagel***, introduced with FiD v2.1, which contains the combination of all ***bioage*** files.

Variables in bioage01-10 and bioagel

The following table provides an overview of all variables contained in the FiD ***bioage*** files in alphabetical order, as well as a comparison with the structure of variables in the SOEP. The variables are shown without their prefix, to increase readability and allow the comparison over the different files.[8] Note that no equivalent to the ***bioage02*** file exists in the SOEP at the time of data collection in 2012. In the SOEP comparison column, an "x" indicates that the variable can be found in the SOEP ***bioagel*** file, too.[9] (Note that the ***bioage01-bioage10*** files in the SOEP have the same data, but other variable names at the moment.) Users interested in combining FiD and SOEP data should use the ***bioagel*** files, which contain all ***bioage*** information on every child. Note that the variables in the ***bioagel*** file do not contain any prefix.

---

[7] Note that since FiD v2.0 from March 2012 we do not name variables, which are similar over the datasets, with additional letters anymore. Variable names such as "MOV5" and "MOV5A" did not prove to be helpful for users, but rather confusing.

[8] Note that in FiDv2.0 from March 2012 as well as in FiDv3.1 from September 2013, several variables in the bioage files have been renamed. In addition, several variable labels have been shortened or changed. In case of questions, please contact the FiD-Team.

[9] Cells with an "(x)" indicate that the variable is included in the SOEP, but either the variable name is different or the information stored does not match. Such differences will be harmonized within the SOEP later. **Users should refer to the variable labels when combining these variables over FiD and SOEP**.

Table 1: Overview variables in ***bioage*** files in FiD and SOEP

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
hhnrakt	Current Wave HH Number	x	x	x	x	x	x	x
sample1	Subsample	x	x	x	x	x	x	
persnr	Never Changing Person ID of child	x	x	x	x	x	x	x
persnrresp	Never Changing Person ID of Mother/Respondent	x	x	x	x	x	x	x
syear	Survey year	x	x	x	x	x	x	x
intid	Interviewer ID	x	x	x	x	x	x	
	VARIABLE PREFIX IS FILE SPECIFIC	**b01**	**b02**	**b03**	**b06**	**b08**	**b10**	
actcare1	Seeking ch. care: Waitinglist	x	x	x				
actcare2	Seeking ch. care: In contact w. nanny/day care	x	x	x				
actcare3	Seeking ch. care: In contact w. welfare office	x	x	x				
actcare4	Seeking ch. care: Other	x	x	x				
activ1	Last 14 ds: Singing children songs		x	x	x			x
activ2	Last 14 ds: Taking walks outdoors		x	x				x
activ3	Last 14 ds: Drawing or doing arts/crafts		x	x	x			x
activ4	Last 14 ds: Reading/telling stories (German)		x	x	x			x
activ5	Last 14 ds: Looking at picture books		x	x				x
activ6	Last 14 ds: Going to playground		x	x	x			x
activ7	Last 14 ds: Visiting oth. families w. children		x	x	x			x
activ8	Last 14 ds: Going shopping w. child		x	x	x			x
activ9	Last 14 ds: Watching TV/videos/DVD w. child		x	x	x			x
activ10	Last 14 ds: Reading/telling stories (not German)		x	x	x			(x)
activ11	Last 14 ds: Actions outside (walks o. similar)				x			(x)
activ12	Last 14 ds: Card games or similar				x			(x)
activ13	Last 14 ds: Playing computer game w. child				x			(x)
activ14	Last 14 ds: Going to theater, circus, museum w. child				x			(x)
adpccare1	Probl. adapting to day care center: Contact with other children	x	x	x	x			
adpccare2	Probl. adapting to day care center: Relationship to child care workers	x	x	x	x			
adpccare3	Probl. adapting to day care center: Separation from home	x	x	x	x			
adpccare4	Probl. adapting to day care center: Adjustment to fixed daily rhythm	x	x	x	x			
adpscl1	Probl. adapting to primary school: Contact with other cildren					x	x	

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
adpscl2	Probl. adapting to primary school: Contact with teachers					x	x	
adpscl3	Probl. adapting to primary school: Want of familiar environment					x	x	
adpscl4	Probl. adapting to primary school: Abiding by school's rules					x	x	
adpscl5	Probl. adapting to primary school: School performance					x	x	
age	Age of child in months	x	x	x	x	x	x	x
allowpm	Child allowance per month					x	x	x
allowpw	Child allowance per week					x	x	x
behav1	Ch. is considerate				x	x	x	x
behav2	Ch. is anxious, overactive, cannot sit quietly				x	x	x	x
behav3	Ch. often complains a. headache/stomach sickness				x	x	x	(x)
behav4	Ch. likes to share with other children				x	x	x	(x)
behav5	Ch. has rage attacks, is choleric				x	x	x	(x)
behav6	Ch. is a loner, plays mostly alone				x	x	x	(x)
behav7	Ch. is compliant, does what adults ask for				x	x	x	(x)
behav8	Ch. has a lot of worries, is depressed				x	x	x	(x)
behav9	Ch. is helpful if anothers are injured/ill/unhappy				x	x	x	(x)
behav10	Ch. is fidgety				x	x	x	(x)
behav11	Ch. has at least one good friend				x	x	x	(x)
behav12	Ch. quarrels with other children, mobs them				x	x	x	(x)
behav13	Ch. is often unhappy or depressed, weeps often				x	x	x	(x)
behav14	Ch. is popular with other children				x	x	x	(x)
behav15	Ch. is easily distracted, unconcentrated				x	x	x	(x)
behav16	Ch. is nervous, clinging in new situations				x	x	x	(x)
behav17	Ch. is nice to younger children				x	x	x	(x)
behav18	Ch. often lies or cheats				x	x	x	(x)
behav19	Ch. is mobbed by others				x	x	x	
behav20	Ch. offers its help voluntary				x	x	x	
behav21	Ch. takes things w/o asking for permission				x	x	x	
behav22	Ch. gets along better w. adults than w. children				x	x	x	
behav23	Ch. has many fears, is easily afraid				x	x	x	
behav24	Ch. finishes tasks, can concentrate for long time				x	x	x	
bepar1	Be parents: sacrifice own wishes					x	x	x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
bepar2	Be parents: disobedience is aimed to bother me					x	x	x
bepar3	Be parents: child upbringing is all that is left					x	x	x
bepar4	Be parents: wish was less captured in responsibility					x	x	x
bepar5	Be parents: problems in upringing is fault of child					x	x	x
bepar6	Be parents: whn with child, nothing is better					x	x	x
bepar7	Be parents: would bear anything for child					x	x	x
bepar8	Be parents: child misbehavior is intentional					x	x	x
bepar9	Be parents: often put everything aside to support child					x	x	x
bepar10	Be parents: look foward to spending time w. child					x	x	x
biochild	Biological child	x	x	x	x			x
biopar	Respondent is mother/father					x	x	
birthm	Month of birth, child	x	x	x	x	x	x	x
birthy	Year of birth, child	x	x	x	x	x	x	x
birthpw	Pregnancy week at birth	x	x	x				x
breastf	Breast-feeding	x	x	x				x
breastfc	Currently breast-feeding	x	x					
breastfm	Breast-feeding time in months	x	x	x				x
care1	Cared for by partner	x	x	x	x	x	x	
care1h	Cared for by partner (hrs/wk)	x	x	x	x	x	x	x
care2	Cared for by moth/fath.(if not in household)		x	x	x	x	x	
care2h	Cared for by moth/fath.(if not in household)(hrs/wk)		x	x	x	x	x	x
care3	Cared for by grandparents	x	x	x	x	x	x	
care3h	Cared for by grandparents (hrs/wk)	x	x	x	x	x	x	x
care4	Cared for by older siblings	x	x	x	x	x	x	
care4h	Cared for by older siblings (hrs/wk)	x	x	x	x	x	x	x
care5	Cared for by other relatives	x	x	x	x	x	x	
care5h	Cared for by other relatives (hrs/wk)	x	x	x	x	x	x	x
care6	Cared for by nanny/day care (not in-house)	x	x	x	x			
care6h	Cared for by babysitter/nanny (hrs/wk)	x	x	x	x			x
care7	Cared for by babysitter/nanny (in-house)		x	x	x	x	x	
care7h	Cared for by babysitter/nanny (in-house)(hrs/wk)		x	x	x	x	x	x
care8	Cared for in crib/day care center	x	x	x	x			

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
care8h	Cared for in crib/day care center (hrs/wk)	x	x	x	x			x
care9	Cared for at school					x	x	
care9h	Cared for at school (hrs/wk)					x	x	x
care10	Cared for in after-school hoard					x	x	
care10h	Cared for in after-school hoard (hrs/wk)					x	x	x
care11	Cared for in social institution, center					x	x	
care11h	Cared for in social institution, center (hrs/wk)					x	x	x
care12	Cared for by others	x	x	x	x	x	x	
care12h	Cared for by others (hrs/wk)	x	x	x	x	x	x	x
care13	No person or institution	x	x	x	x	x	x	x
careaft	Child in school: Problems with afternoon child care				x			
careaftw	Child n.i. school: Worries about afternoon child care				x			
caremn	Start current child care, month				x			
caretyp	Type of child care, if any				x			
careyr	Start current child care, year				x			
ccare	Child care: day care center/nanny (y/n)	x	x	x				
ccarec	Never changed child care institution		x	x	x			
ccarecn	Number of changes of child care institution		x	x	x			
ccarefr	Child care: freq of day care center/nanny	x	x	x				
change1	Life circumstances have greatly changed	x	x	x				x
change2	Child provides happiness and joy	x	x	x				x
change3	Often close to running out of strength	x	x	x				x
change4	Very satisfied with the role of (being) a mother	x	x	x				x
change5	Often unable to cope with (new) tasks/responsibil.	x	x	x				x
change6	Have made new contacts through the child	x	x	x				x
change7	Suffer f. being limited to role of mother/father	x	x	x				x
change8	Important to provide the child w. much affection	x	x	x				x
char1	Ch. is communicative vs. quiet				x		x	(x)
char2	Ch. is untidy vs. tidy				x		x	(x)
char3	Ch. is sweet-tempered vs. short-tempered				x		x	(x)
char4	Ch. is not interested vs. eager to learn				x		x	(x)
char5	Ch. has self-confidence vs. is insecure				x		x	(x)

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
char6	Ch. is withdrawn vs. sociable				x		x	(x)
char6y	Child tends to be timid vs. sociable		x	x				(x)
char7	Ch. is concentrated vs. easily distracted				x		x	(x)
char7y	Child tends to be concentrated vs. distracted		x	x				(x)
char8	Ch. is defiant vs. obedient				x		x	(x)
char8y	Child tends to be defiant vs. obedient		x	x				(x)
char9	Ch. comprehends fast vs. needs more time				x		x	(x)
char9y	Child tends to comprehend fast vs. needs more time		x	x				(x)
char10	Ch. is afraid vs. unafraid				x		x	x
chhealth	Child health status					x	x	x
circum	Head circumference of the child at birth in cm	x	x	x				x
clofrie	Number of close friends						x	
conscho1	Contact w. school: Reg. partc. at parent-teacher evening						x	x
conscho2	Contact w. school: Reg. parent-teacher meeting						x	x
conscho3	Contact w. school: See teacher individually						x	x
conscho4	Contact w. school: Parent representative						x	x
conscho7	Contact w. school: None of these						x	x
curscol1	Cur. school: Primary School					x	x	x
curscol2	Cur. school: Special educational concepts					x	x	x
curscol3	Cur. school: Special needs school					x	x	x
curscol4	Cur. school: 'Hauptschule'						x	x
curscol5	Cur. school: 'Realschule'						x	x
curscol6	Cur. school: 'Gymnasium'						x	x
curscol7	Cur. school: 'Gesamtschule'						x	x
curscol8	Cur. school: Other school					x	x	x
delivcs	Delivery by caesarean section	x	x	x				x
delivpl	Place of delivery	x	x	x				x
dint	Day of interview	x	x	x	x	x	x	
dis0	Child is restricted in abilities				x			
disord	Child has confirmed disorders	x	x	x				x
disord1	Confirm. disorders: Sensory (vision&hearing)	x						
disord2	Confirm. disorders: Motor functions	x						

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
disord3	Confirm. disorders: Neurological	x						
disord4	Confirm. disorders: Speech	x						
disord5	Confirm. disorders: Regulatory	x						
disord6	Confirm. disorders: Chronic disease	x						
disord7	Confirm. disorders: Physical disability	x						
disord8	Confirm. disorders: Mental disability	x						
disord9	Confirm. disorders: Other	x						
disordu	First U exam confirming disorder		x	x				
dvlpm1	Problems in development/behavioral problems				x			
dvlpm2	In therapy/consulting because of behav./devel. problems				x			
edbeh1	Edu. behav.: show love					x	x	x
edbeh2	Edu. behav.: criticize					x	x	x
edbeh3	Edu. behav.: ask what he/she has experienced					x	x	x
edbeh4	Edu. behav.: punish when disobedient					x	x	x
edbeh5	Edu. behav.: threaten to punish, but do not punish					x	x	x
edbeh6	Edu. behav.: know where child is					x	x	x
edbeh7	Edu. behav.: rather strict with child					x	x	x
edbeh8	Edu. behav.: comfort child when child is sad					x	x	x
edbeh9	Edu. behav.: shout when child makes mistakes					x	x	x
edbeh10	Edu. behav.: child is not grateful, b/c is disobedient					x	x	x
edbeh11	Edu. behav.: do not talk to child when disobedient					x	x	x
edbeh12	Edu. behav.: tell child to not disobey					x	x	x
edbeh13	Edu. behav.: praise my child					x	x	x
edbeh14	Edu. behav.: insult child b/c I am upset w. him/her					x	x	x
edbeh15	Edu. behav.: try to influence childs friendships					x	x	x
edbeh16	Edu. behav.: reduce punishmnt or cancel it					x	x	x
edbeh17	Edu. behav.: disappointed about bad behavior					x	x	x
edbeh18	Edu. behav.: hard to be consistent in upbringing					x	x	x
edgoal1	Child should: be a good student					x	x	x
edgoal2	Child should: get along with other kids					x	x	x
edgoal3	Child should: be interested why things happen					x	x	x
edgoal4	Child should: act like normal girl/boy					x	x	x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
edgoal5	Child should: be honest					x	x	x
edgoal6	Child should: have good manners					x	x	x
edgoal7	Child should: have good self-control					x	x	x
edgoal8	Child should: be responsible					x	x	x
edgoal9	Child should: treat others with respect					x	x	x
edgoal10	Child should: obey the parents					x	x	x
edgoal11	Child should: have good ability to judge					x	x	x
edgoal12	Child should: be clean and neat					x	x	x
edgoal13	Child should: strive to achieve aims					x	x	x
edgoal14	Child should: fit in well in groups					x	x	x
edgoal15	Child should: learn to fight obstacles					x	x	x
edgoal16	Child should: be satisfied with self					x	x	x
edgoal17	Child should: learn to avoid risks					x	x	x
edgoal18	Child should: be lovable					x	x	x
famofreq	Freq. child sees mother/father		x	x	x			
famoinhh	Bio. fath./moth. (of child) lives in househ.		x	x	x			
fathfreq	Freq. child sees father	x						
fathinhh	Bio. father of child lives in household	x						x
feeling1	Physic. condition in final trimester	x						x
feeling2	Physic. condition in 1st 3 months after birth	x						x
feeling3	Mental state in final trimester	x						x
feeling4	Mental state in 1st 3 months after birth	x						x
freqact1	Freq. of activities: Watching tv, video, dvd						x	x
freqact2	Freq. of activities: Computer games						x	x
freqact3	Freq. of activities: Surfing the web/chatting						x	x
freqact4	Freq. of activities: Listening to music						x	x
freqact5	Freq. of activities: Playing music						x	x
freqact6	Freq. of activities: Doing sports						x	x
freqact7	Freq. of activities: Doing sth. w. family						x	x
freqact8	Freq. of activities: Dancing, theater						x	x
freqact9	Freq. of activities: Metal-/woodwork						x	x
freqact10	Freq. of activities: Painting/handicraft						x	x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
freqact11	Freq. of activities: Reading						x	x
freqact12	Freq. of activities: Hanging out, dreaming						x	x
freqact13	Freq. of activities: Being together with friends						x	x
freqact14	Freq. of activities: Going to church						x	x
health	Concerns about the child health	x	x	x				x
height	Height of child in cm		x	x	x			x
heightb	Height of child at birth in cm	x	x	x				
helphmwk1	Frequency help with homework by mother					x		
helphmwk2	Frequency help with homework by father					x		
helphmwk3	Frequency help with homework by siblings					x		
helphmwk4	Frequency help with homework by friends					x		
helphmwk5	Frequency help with homework by tutorv					x		
helphmwk6	Frequency help with homework by homework supervision					x		
sathmwk	Satisfaction help with homework					x		
hospital3mb	Days in hospital 1st 3 months after birth	x	x	x				x
hospital12m	Days in hospital last 12 monhts		x	x	x	x	x	x
idegrad1	Ideal grad: low ('Hauptschule')					x	x	x
idegrad2	Ideal grad: medium ('Realschule')					x	x	x
idegrad3	Ideal grad: high ('Abitur')					x	x	x
ill2	Disease: Otitis media		x	x	x	x		x
ill3	Disease: Hay fever		x	x				(x)
ill4	Disease: Neurodermatitis		x	x	x	x		x
ill5	Disease: Ametropia		x	x	x	x		x
ill6	Disease: Hardness of hearing		x	x				x
ill7	Disease: Nutritional disturbances		x	x	x	x		x
ill8	Disease: Disturbance of motoric functions		x	x	x	x		x
ill9	Disease: Other		x	x	x	x		x
ill10	Disease: respiratory disease				x	x		x
ill11	Disease: Asthma		x	x				x
ill12	Disease: Chronic bronchitis		x	x				x
ill13	Disease: Akute bronchitis		x	x				x
ill14	Disease: Croup syndrome		x	x				x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
illno	No confirmed disease		x	x	x	x		x
imper1a	1st person: child/adult						x	
imper1b	1st person: relative/not related						x	
imper1c	1st person: gender						x	
imper1d	1st person: age						x	
imper1e	1st person: German origin						x	
imper1f	1st person: education						x	
imper2a	2nd person: child/adult						x	
imper2b	2nd person: relative/not related						x	
imper2c	2nd person: gender						x	
imper2d	2nd person: age						x	
imper2e	2nd person: German origin						x	
imper2f	2nd person: education						x	
imper3a	3rd person: child/adult						x	
imper3b	3rd person: relative/not related						x	
imper3c	3rd person: gender						x	
imper3d	3rd person: age						x	
imper3e	3rd person: German origin						x	
imper3f	3rd person: education						x	
inshelp1	Used help: family education facility						x	
inshelp2	Used help: family counseling						x	
inshelp3	Used help: support by youth-welfare-center						x	
inshelp4	Used help: support services by social workers						x	
inshelp5	Used help: day-cares services by social workers						x	
inshelp6	Used help: consultation of nurse/educator						x	
inshelp7	Used help: school-counselor/therapist						x	
inshelp8	Used help: other aid of youth-welfare-center						x	
inshelp9	Used help: other aid						x	
lamark	Mark/grade: German					x	x	x
language	Language spoken with child at home	x	x	x	x	x	x	x
language1	Language 1	x	x	x	x	x	x	
language2	Language 2	x	x	x	x	x	x	

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
language3	Language 3	x	x	x	x		x	
lstmedex	Last med. examination (U-exam)	x	x	x				x
maincare	Respondent(mother) is main child care provider	x				x	x	x
matmark	Mark/grade: Math					x	x	x
medaid3m	Medical help: # of times in last 3 months		x	x	x	x	x	x
medaid3mb	Medical help: # of times 1st 3 months after birth	x	x	x				x
mint	Month of interview	x	x	x	x	x	x	
mvmn1	Child walks down stairs forward			x				x
mvmn2	Child uses door handle to open doors		x	x				x
mvmn3	Child uses climbing frames + high playground equip.			x				x
mvmn4	Child uses scissors to cut paper			x				x
mvmn5	Child paints/draws recognizable forms on paper			x				x
mvmn6	Child jumps with one foot			x				
mvmn7	Child holds pens correctly		x	x				(x)
mvmn8	Child climbs stairs w. alternating feet		x					
mvmn9	Child can run w/o falling down		x					
mvmn10	Child is able to jump forwards at least 3 times		x					
mvmn11	Child can color simple forms and figures		x					
mvmn12	Child can solve puzzles w. a min. of 2 parts		x					
nactcar1	Not seek. child care: but concrete future plans	x	x	x				
nactcar2	Not seek. child care: take it as it comes	x	x	x				
nactcar3	Not seek. child care: general disapproval of ch. care	x	x	x				
nchild	Child birth order	x	x	x	x			x
noccare1	Reason no care: Child too young	x	x	x	x			
noccare2	Reason no care: Want to raise child by myself	x	x	x	x			
noccare3	Reason no care: At home anyway and can take care	x	x	x	x			
noccare4	Reason no care: Costs are too high	x	x	x	x			
noccare5	Reason no care: No spots available	x	x	x	x			
noccare6	Reason no care: Distance too far	x	x	x	x			
noccare7	Reason no care: Opening hours not suitable	x	x	x	x			
noccare8	Reason no care: Transfers too time-consuming	x	x	x	x			
noccare9	Reason no care: Child has a chronic disease/disorder	x	x	x	x			

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
noccare10	Reason no care: Child should spend time w. siblings	x	x	x	x			
noccare11	Reason no care: Good experiences caring at home				x			
prebeg	Spell Begin Pregnancy (Month, 01.83=1)	x						x
preend	Spell End Pregnancy, Birth (Month,01.83=1)	x						x
pregmo	Mother: pregnancy month at interview	x						x
pregplan	Pregnancy unintended/intended	x	x	x				x
pregy	Mother: pregnant at interview, year	x						x
probgra1	Prob. of grad: low ('Hauptschule')					x	x	x
probgra2	Prob. of grad: medium ('Realschule')					x	x	x
probgra3	Prob. of grad: high ('Abitur')					x	x	x
reccare1	Reason child care: Employment	x	x	x	x			
reccare2	Reason child care: Increased work schedule	x	x	x	x			
reccare3	Reason child care: start/contin. acadm. studies	x	x	x	x			
reccare4	Reason child care: Positive impact on child's develop.	x	x	x	x			
reccare5	Reason child care: Want more time for myself	x	x	x	x			
reccare6	Reason child care: Other	x	x	x	x			
relapare	Relationship to biological mother/father	x	x	x	x			
saccare1	Satisf. child care: with group size	x	x	x	x			
saccare2	Satisf. child care: with size of staff	x	x	x	x			
saccare3	Satisf. child care: with opening hours	x	x	x	x			
saccare4	Satisf. child care: with costs	x	x	x	x			
saccare5	Satisf. child care: with flexibility	x	x	x	x			
saccare6	Satisf. child care: with conduct of staff	x	x	x	x			
saccare7	Satisf. child care: with activities/education	x	x	x	x			
saccare8	Satisf. child care: with parental participation		x	x	x			
saccare9	Satisf. child care: with contact to nurse/teacher		x	x	x			
saccare10	Satisf. child care: with preparation for school				x			
saticare	General satisfaction with child care	x	x	x	x	x	x	
scactiv1	Other school activitiy: athletics						x	
scactiv2	Other school activitiy: acting/dancing						x	
scactiv3	Other school activitiy: choir/orchestra/music group						x	
scactiv4	Other school activitiy: other						x	

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
sclenrolm	Start of school (ever), month					x	x	x
sclenroln	Not yet in school					x		x
sclenroly	Start of school (ever), year					x	x	x
sclr1	Child calls familiar people by name			x				x
sclr2	Child plays games with other Children			x				x
sclr3	Child participates in role-playing ('as if')			x				x
sclr4	Child shows particular liking for certain friends			x				x
sclr5	Child specif. own feelings, e.g. sad, happy			x				x
sclr6	Child is happy or worried for others			x				
sclr7	Child takes turns when playing w/o being asked			x				(x)
sclr8	Child plays games with other Children		x					
sclr9	Child shows desire to make others feel happy		x					
sclr10	Child shares toys if asked for		x					
sclr11	Child continues playing w/o crying whn parents leave		x					
sclr12	Child creatively plays with household stuff		x					
sclr13	Child seeks to make friends with same-age kids		x					
sclr14	Child fulfills simple requests		x					
scolcon1	Schl. cond.: likes to go to school					x	x	x
scolcon2	Schl. cond.: does not get along with classmates					x	x	x
scolcon3	Schl. cond.: thinks school is waste of time					x	x	x
scolcon4	Schl. cond.: does not take school work seriously					x	x	x
scolcon5	Schl. cond.: follows lessons well					x	x	x
scolcon6	Schl. cond.: does not get along w. teacher					x	x	x
scolcon7	Schl. cond.: likes to study/has a zeal for learning					x	x	x
scolcon8	Schl. cond.: gets along w. classmates (SC 2010 only)					x	x	
scolcon9	Schl. cond.: gets along w. teacher (SC 2010 only)					x	x	
scoldura	Time in school in months					x	x	
sex	Gender of child	x	x	x	x	x	x	x
skll1	Child eats w. spoon w/o spilling		x	x				x
skll2	Child blows nose w/o assistance		x	x				x
skll3	Child uses toilet to do number two		x	x				x
skll4	Child puts on pants and underpants frontwards			x				x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
skll5	Child brushes teeth w/o assistance			x				x
skll6	Child closes press buttons w/o assistance		x	x				
skll7	Child covers mouth and nose when sneezing			x				
skll8	Child takes off sweater/jacket w/o assistance		x					
skll9	Child drinks f. a cup/glass w/o spilling		x					
skll10	Child washes and dries its face w/o assistance		x					
spch1	Child understands brief instructions			x				x
spch2	Child forms sentences with at least two words			x				x
spch3	Child speaks in full sentences (at least four words)			x				x
spch4	Child attends orders given by someone 5 minutes ago			x				(x)
spch5	When asked, child gives first and last name			x				(x)
spch6	Child is able to inform others w. simple messages			x				(x)
spch7	Child listens attentively to story f. at least 15 min.			x				(x)
spch8	Child follows brief instructions w. simple tasks		x					
spch9	Child listens attentively to story f. at least 5 min.		x					(x)
spch10	Child can identify at least 5 body parts		x					
spch11	Child can identfy daily-life-objects in a book		x					
spch12	Child answers questions in words, or tries to		x					
spch13	Child can name at least 10 objects		x					
spch14	Child can describe things in simple words		x					
specned1	Special needs: learning					x	x	
specned2	Special needs: linguistic					x	x	
specned3	Special needs: behavior/experience					x	x	
specned4	Special needs: mental					x	x	
specned5	Special needs: amplyopia (eyes)					x	x	
specned6	Special needs: hearing					x	x	
specned7	Special needs: physical/motoric					x	x	
specned8	Special needs: long-lasting disease					x	x	
specned9	Special needs: none					x	x	
stcaremn	Start of child care (ever), month	x	x	x	x			
stcareyr	Start of child care (ever), year	x	x	x	x			
suppartn	Feel supported by partner	x						x

Name	Label	Bioage01	Bioage02	Bioage03	Bioage06	Bioage08	Bioage10	SOEP
temp1	Child generally happy and satisfied	x	x	x				x
temp2	Child is easily irritated, cries often	x	x	x				x
temp3	Child hard to console	x	x	x				x
temp4	Child is curious and active	x	x	x				x
temp5	Child tends to be shy	x						x
temp6	Child likes to talk			x				x
temp7	Child shows empathy when others are sad		x	x				x
timfamod	Time father/mother sees child (days)		x	x	x			
timfamoh	Time father/mother sees child (hours)		x	x	x			
timfathd	Time father sees child (days)	x						
timfathh	Time father sees child (hours)	x						
tvhrs	Watches video/tv alone (h/week)			x	x			x
tvyn	Child may watch tv unattended			x	x			x
vistfrnd	Time child spends at friends (alone)			x	x			
weight	Weight of child in kilograms		x	x	x			x
weightb	Weight of child at birth in grams	x	x	x				x

Topics covered in the bioage files

The ***bioage*** data files cover 18 different subjects regarding various aspects of children's upbringing. The following table provides an overview of which subject is covered in which file. For more specific information please refer to the parent-questionnaires.

Table 2: Overview of topics

Subejcts covered	**Bioage File:**					
	01	**02**	**03**	**06**	**08**	**10**
Age group (years)	**0-1**	**1-2**	**2-3**	**5-6**	**7-8**	**9-10**
Information regarding child birth	**x**	**x**	**x**			
Physical and mental well-being during pregnancy and at the time of giving birth	**x**					
Changed circumstances and first experiences of living with a child	**x**	**x**	**x**			
Child's health, medical examinations and medical treatment	**x**	**x**	**x**	**x**	**x**	**x**
Relationship to biological mother/father	**x**	**x**	**x**	**x**		
Activities done with the child		**x**	**x**	**x**		
Leisure-time						**x**
Child care	**x**	**x**	**x**	**x**	**x**	**x**
School					**x**	**x**
Peer groups				**x**	**x**	**x**
Child's network						**x**
Child's temper and personality	**x**	**x**	**x**	**x**	**x**	**x**
Child's adaptive behavior		**x**	**x**			
Child's socio-emotional behavior		**x**	**x**	**x**	**x**	**x**
Language spoken in household	**x**	**x**	**x**	**x**	**x**	**x**
Allowance					**x**	**x**
Parental ambitions regarding the education of the child					**x**	**x**
Being parents					**x**	**x**
Upbringing: preferences/principles					**x**	**x**
Educational aims					**x**	**x**

Data files and respondents

Data files are named according to the age of a child found within a specific ***bioage*** data file, corresponding with the file specific suffix (e.g. ***bioage06*** includes children turning six in the survey year, giving a range between 5 years and 1 month and 6 years and 11 months, depending on month of birth and interview month). For the correspondence between age groups and questionnaires please consult table 3. Whether the mother or the father or both answer the questionnaire and thus are the (proxy) respondent for the ***bioage*** data depends on the respective questionnaire.

***bioage01*: ‘Parent-Questionnaire 1: 0-1 year old’**

The questionnaire is given to all women who gave birth to their own child in the current survey year or the year before. In addition, all women whose non-biological child was born in this time period receive the questionnaire. In contrast to parent-questionnaires for older children, because of some of the contents, the “Parent-Questionnaire 1” is meant to be answered by the mother (in a very few cases fathers have answered the questionnaire).

***bioage02*: ‘Parent-Questionnaire 2: 1-2 years old’**

Principally, this questionnaire is given to all mothers whose child turns two in the current survey year (this applies to biological, as well as non-biological children). Note, however that in case the father is a single parent he may answer the questionnaire as well.

***bioage03*: ‘Parent-Questionnaire 3: 2-3 years old’**

The questionnaire is given to all parents whose child turns three in the current survey year. It can be answered either by the mother or father.

***bioage06*: ‘Parent-Questionnaire 4: 5-6 years old’**

The questionnaire is given to all parents with children turning six years old during the survey year. It should be answered either by the mother or by the father.

***bioage08*: ‘Parent-Questionnaire 5: 7-8 years old’**

The questionnaire is given to *both* mothers and fathers of children turning eight years old in the current survey year. Two questionnaires will be filled out only if either (biological or non-biological) parents live in the household.

bioage10: 'Parent-Questionnaire 6: 9-10 years old'

The questionnaire is given to *both* mothers and fathers of children turning eight years old in the current survey year. Two questionnaires will be filled out only if either (biological or non-biological) parents live in the household.

Table 3: Overview questionnaire names and corresponding age group

Bioage File	Classification by Age	Quest. Field Name Wave 2010	Quest. Field Name Wave 2011 and later	Respondents
bioage01	Parent-Quest. 0-1 years old	Parent-Quest.1A	Parent-Quest. 1	Mother only
bioage02	Parent-Quest. 1-2 years old	Parent-Quest.1B	Parent-Quest. 2	Mother or father
bioage03	Parent-Quest. 2-3 years old	Parent-Quest. 2	Parent-Quest. 3	Mother or father
bioage06	Parent-Quest. 5-6 years old	Parent-Quest. 3	Parent-Quest. 4	Mother or father
bioage08	Parent-Quest. 7-8 years old	Parent-Quest. 4	Parent-Quest. 5	Mother and father
bioage10	Parent-Quest. 9-10 years old	Parent-Quest. 5	Parent-Quest. 6	Mother and father

Specifics of *bioage08* & *bioage10*

Due to the fact that parent-questionnaires 5 and 6 are given to both parents of the child, the dataset of ***bioage08*** and ***bioage10*** was split into ***bioage08p1/bioage10p1*** for the first respondent and ***bioage08p2/bioage10p2*** for the second respondent. Typically, the first respondent is the biological mother or a single parent. Thus, for each child there exists a second observation provided the second parent filled out the respective questionnaire. The actual number of children is congruent with the number of observations in ***bioage08p1/bioage10p1***, while the number of observations in ***bioage08p2/bioage10p2*** are those observations, which are additionally available for the children of that age group (note that this apllies only to children for whom two questionnaires have been filled out).

Table 4: Overview number of observations in ***bioage08 p1*** and ***p2***

	Bioage08p1	***Bioage08p2***	***Bioage08* (total)**
Mother	1809	0	1809
Father	55	1276	1331
Total	1864	1276	3140

Table 5: Overview number of observations in ***bioage10 p1*** and ***p2***

	Bioage10p1	***Bioage10p2***	***Bioage10* (total)**
Mother	1782	0	1782
Father	67	1144	1211
Total	1849	1144	2993

***bioagel*: 'Combined dataset of all bioage files'**

Starting with the data distribution FiD v2.1, a combined dataset of all **bioage** files is distributed, analogously to the SOEP dataset first distributed with SOEP v28. This file is called ***bioagel***, where the "*l*" characterized the dataset as being in the "long" format, i.e. multiple observations per child are present in the dataset, where variables do not have a prefix anymore, but are identically named in case the question across the parent questionnaires are identical[10]. The ***bioagel*** dataset allows following children and their developments within one dataset: e.g., one can now see how children have grown from one interview to the next. The data are organized on a child-dataset level, i.e. the variables PERSNR and BIOAGE (see below) identify unique observations. Note that any variable not asked because the child was not in the respective age group will receive the missing code "-5 Question not included". The ***bioagel*** dataset will become more and more useful with every year of additional observations.

[10] Note that from FiD v3.0 the variables B10BEHAV in ***bioage10*** are no longer an exception to this rule and all these variables are presented in a three-point scale (see below).

Number of observations

Table 6: Overview number of observations per wave

	Number of observations				
Filename	**2010**	**2011**	**2012**	**2013**	**Total**
bioage01	1,321	207	212	167	1,907
bioage02	787	647	568	187	2,189
bioage03	871	741	555	523	2,690
bioage06	473	486	425	656	2,040
bioage08p1	425	529	501	409	1,864
bioage08p2	257	373	348	298	1,276
bioage10p1	403	510	477	459	1,849
bioage10p2	242	310	291	301	1,144
Total	4,779	3,803	3,377	3,000	14,959

Generated Variables

This section provides additional information on variables, which have been generated from combinations of variables from other datasets than the parent questionnaires.

AGE

Variable label **"Child's age in months"**
Variable format 2-digit integer
Bioage File 01-10

Comment: The variable AGE provides the child's age in months, as a combination of month of birth information and the interview month. As the exact day of birth remains unknown, information is only an approximation and may vary by one month.

BREASTFM

Variable label **"Breast-feeding time in months"**
Variable format 2-digit integer
Bioage File 01; 02; 03

Comment — The breast-feeding time in months is taken directly from the questionnaire, if the child is no longer breast-fed and the mother stated the number of months. If the baby is currently breast-fed, BREASTFM is set to "-2 Does not apply", because the information is not identical. However, assuming that there were no gaps in breast-feeding, users can easily generate BREASTFM for children who are currently breast-fed by taking the age of the child in months (AGE).

PREBEG

Variable label — **"Spell Begin Pregnancy (Month, 01.83=1)"**
Variable format — 3-digit integer
Bioage File — 01

Comment — This variable is based on the exact month of birth (BIRTHM) and the duration of childbearing in weeks (BIRTHPW). Accordingly, information is available only for women who completed the "Parent-Questionnaire 1" and for whom the duration of the pregnancy is known. Note that the month of conception may vary by one month, as the exact *day* of birth remains unknown.
The variable PREBEG contains information on the beginning of childbearing (e.g. the month of conception). Information in FiD is given in the regular SOEP spell format: values start with 1 for January 1983 (e.g. earliest spell in ***bioage01***, survey year 2010, is 304, which equals April 2008). This spell design allows for direct comparison with spell information in the SOEP data files (e.g. ***biomarsm***, ***artkalen***).

PREEND

Variable label — **"Spell End Pregnancy, Birth (Month, 01.83=1)"**
Variable format — 3-digit integer
Bioage File — 01

Comment — This variable is based on the exact month of birth (BIRTHM) and the duration of childbearing in weeks (BIRTHPW). Accordingly, information is available only for women who completed the "Parent-Questionnaire 1" and for whom the duration of the pregnancy is known.
Variable PREEND contains information on the end of childbearing (e.g. the month of birth). Information in FiD is given in the regular SOEP spell format: values start with 1 for January 1983 (e.g. earliest spell in ***bioage01***, survey year 2010, is 304, which equals April 2008). This spell design allows for direct comparison with spell information in the SOEP data files (e.g. ***biomarsm***).

PREGY

Variable label — **"Mother: pregnant at interview, year"**
Variable format — 4-digit integer
Bioage File — 01

Comment — This variable is based on information the mother gave on her then current pregnancy status in the previous year in the person-questionnaire. In case a pregnancy is stated (or was unknown) and a child was born, the year in which the interview was done is recorded in PREGY. Hence this information is available only for those women in the sample for at least two years with completed personal interview in the first and a completed parent-questionnaire 1 in the second year. Please note that future mother need not be aware of their pregnancy in the early stages of childbearing.

PREGMO

Variable label — **"Mother: pregnancy month at interview"**
Variable format — 2-digit integer
Bioage File — 01

Comment — This variable is based on the exact month of birth (BIRTHM), the duration of childbearing in weeks (BIRTHPW) and the interview month of the previous year's personal interview. Hence this information is available only for those women in the sample for at least two years. As the exact day of birth is unknown, this variable remains a close estimation.

SEX

Variable label — **"Gender of child"**
Variable format — 1-digit integer
Bioage File — 01-10

Comment — The sex of the child has not been asked for in the Parent-Questionnaires. Information for this variable stems from the ***ppfad***.

SEXRESP

Variable label — **"Gender of respondent (parent)"**
Variable format — 1-digit integer
Bioage File — 01-10

Comment — This variable can be used to identify whether the mother or father of the child has answered the respective questionnaire. The information for this variable comes from the ***ppfad*** file.

CARE6

Variable label — **"Cared for by nanny/day care (not in-house)"**
Variable format — 1-digit integer
Bioage File — 01; 02; 03

Comment — Within the parent questionnaires 1-3 there are some discrepancies when comparing the first question on child care (related to day care and nannies only, variables $E121b, $E224, $E324) with the second one

(dealing with child care in general, variables $E128, $E232, $E332). Specifically, instances of nannies (not in-house) and day care centers were reported less in the latter questions, which also does not correspond with information given in the household questionnaire (dataset ***$kind***, variables $K065A, $K065B, $K066A, $K066B, $K070D). We cannot be sure of the cause of these differences, we assume, however that it was not clear to some respondents why they were meant to give partially identical information in the two questions in the parent questionnaire. This assumption is based on the CAPI interviews, where the differences are more pronounced. In CAPI mode, first the list of care givers is asked in questions $E128, $E232, and $E332, making the items almost identical to the first question. Only then the respondents are asked to provide the number of hours each care giver spends with the child. In PAPI mode, the respondents immediately see that they are asked to provide the hours as well, and hence understand the purpose of the question. When the person filling out the household questionnaire is identical to the one doing the parent questionnaire, the problem is also slightly more present. We interpret this again in the direction that individuals do not understand why they should give the same information twice (or even three times).

These problems are tackled by combining all information available for nannies (CARE6) and day care centers (CARE8) as well as hours spent at day care centers (CARE8H). We use the household information for each child (from ***$kind***) together with the information from the respective parent questionnaire to provide more reliable measures. The most reliable information is assumed to be the household questionnaire information – hence hours and incidence are taken from here if there is information available. Only when the parent questionnaire provides additional information, it is used and coded in the respective variable.

Of course the solution we provide here is only one of many. As all information from parent and household questionnaires are available to the user, other ways to deal with these issues are possible. However, we advise to use caution when using the variables $E128E1, $E128E2, $E128F1, and $E128F2 in parent questionnaire 1, $E232F1, $E232F2, $E232H1, and $E232H2 in parent questionnaire 2, and $E332F1, $E332F2, $E332H1, and $E332H2 in parent questionnaire 3.

To cope with the problems mentioned above, the questionnaire was slightly changed in 2013, where the question on nannies and dare care centers was asked only once (e.g. questions 21 and 29 in parent questionnaire 1). In addition to whether the child was cared for by a nanny or day care center, information on the duration of care per day was collected. In the later question (where information on other sources of care is collected) day care and nannies are not included any more. This led to two changes in the coding of the variable: first, CARE6 and CARE8 had to be generated via CCAREFR (frequency of child care). Second, as CARE6H and CARE8H are now captured in hours per day, the values were multiplied by five to adapt this information to the structure of the previous waves (hours per week).

CARE6H

Variable label	**"Cared for by nanny/day care (not in-house) (hrs/wk)"**
Variable format	3-digit integer
Bioage File	01; 02; 03

Comment: See comment for CARE6

CARE8

Variable label	**"Cared for in crib/day care center"**
Variable format	1-digit integer
Bioage File	01; 02; 03

Comment See comment for CARE6

CARE8H

Variable label	**"Cared for in crib/day care center (hrs/wk)"**
Variable format	3-digit integer
Bioage File	01; 02; 03

Comment: See comment for CARE6

CARE9H – CARE11H

Variable label	**"Cared for at school/ in after-school hoard/ in social institution, center (hrs/wk)"**
Variable format	3-digit integer
Bioage File	08; 10

Comment: In 2013 the information on CARE9H, CARE10H and CARE11H was captured in terms of hours per day. To adapt this information to the structure of the previous waves (hours per week) we multiplied it by five. However, this created some inconsistencies which probably derive from the nature of the question. Since the previous questions were asked in terms of hours per week, it seemed that some respondents still answered the CARE9H – CARE11H for the whole week. Therefore the answers were only multiplied by five when they did not exceed 12 hours, assuming that no child spends more than 12 hours a day in care facilities and that other answers were meant to reflect the weekly hours. Thus, answers over 12 remained unchanged.

CARE1H - CARE13H

Variable Label	**"Cared for by <care giver> (hrs/wk)"**
Variable format	3-digit integer
Bioage File	01; 02; 03; 06; 08; 10

Comment For all positive hours given for the different care givers, we sum them up and check whether they are larger than 168 (the maximum hours per

week). If so, we reduce each positive number of hours according to the percentage they make up from the total and round them to the nearest full hour. This is done after any corrections are applied to CARE8H.

BIOAGE

Variable label	**"Original bioage dataset"**
Value labels	BIOAGE (1) Bioage01 (2) Bioage02 (3) Bioage03 (6) Bioage06 (81) Bioage08p1 (82) Bioage08p2 (101) Bioage10p1 (102) Bioage10p2
Variable format	3-digit integer
Bioage File	***bioagel***
Comment	The variable BIOAGE is only given in the long version of the ***bioage*** data, in ***bioagel***. It allows identifying the source of the information, i.e. from which ***bioage*** dataset a variable was drawn. For a distribution of BIOAGE, refer to table 6.

Different questionnaire versions

While the main part of the questionnaires has remained constant across the different samples and across the two waves of data collection, there have been a few changes, which are documented here.

bioage02

Information on birth circumstances for children of 2010, who were already covered by the parent-questionnaires 1, is no longer collected in the following waves. To be able to separate this (purely retrospective) part visually for the respondents, to some small changes in question ordering were needed (see table 7).

Table 7: Changes in question ordering, retrospective and non-retrospective birth questions

Parent questionnaire 2	**2010 Cohort Sample**	**2011 and 2012 (all samples)**
Planned Pregnancy	Question 6	Question 6, retrospective
Place of delivery	Question 7	Question 7, retrospective part
Pregnancy week of delivery	Question 8	Question 8, retrospective part
Height, weight and head circumference at birth	Question 9	Question 9, retrospective part
Breast-feeding	Question 10	Question 12, not in retrospective part
Disorders at medical exams	Question 11	Question 13, not in retrospective part
Last U-exam	Question 12	Question 14, not in retrospective part
Medical help needed during 1st 3 months after birth	Question 13	Question 10, retrospective part
Hospital visits during 1st 3 months after birth	Question 14	Question 11, retrospective part

Note also that the parent-questionnaire 2 was fielded by mail for the screening sample in 2010 after the data collection had already finished (it is the only questionnaire done by mail in FiD).

bioage03

The parent-questionnaire 3 in the screening sample of 2010 lacked information on birth circumstances of the child. This was later added to the questionnaire in the cohort sample in 2010, and is now a fixed part of this questionnaire. In addition, similar to the parent-

questionnaire 2, the retrospective information on birth circumstances is not asked a second time. These changes bring about new information as well as changes in question ordering documented in table 8. Note that information missing due to questions not being asked is coded "[-2] Does not apply".

Table 8: Additions to questions and changes in order of questions in parent questionnaire 3

Parent questionnaire 3	2010 Screening Sample	2010 Cohort Sample	2011/2012 (all samples)
Child's birth order	Not included	Question 3	Question 3, not in retrospective part
Biological child	Not included	Question 4	Question 4, not in retrospective part
Changes after birth	Not included	Question 5	Question 5, not in retrospective part
Planned Pregnancy	Not included	Question 6	Question 6, retrospective
Place of delivery	Not included	Question 7	Question 7, retrospective
Pregnancy week of delivery	Not included	Question 8	Question 8, retrospective
Height, weight and head circumference at birth	Not included	Question 9	Question 9, retrospective
Breast-feeding	Question 3	Question 10	Question 12, not in retrospective part
Disorders at medical exams	Not included	Question 11	Question 13, not in retrospective part
Last U-exam	Not included	Question 12	Question 14, not in retrospective part
Medical help needed during 1st 3 months after birth	Not included	Question 13	Question 10, retrospective
Hospital visits during 1st 3 months after birth	Not included	Question 14	Question 11, retrospective
Current height and weight	Question 7	Question 15	Question 15
Hospital visits in last 12 months	Question 4	Question 16	Question 16
Medical help needed in last 3 months	Question 5	Question 17	Question 17
Illnesses or disorders confirmed by doctor	Question 6	Question 18	Question 18
Rest of the questionnaire corresponds one-to-one	Question 8 through question 28	Question 19 through question 39	Question 19 through question 39

bioage06

The parent-questionnaire 4 for children aged five to six the information on birth order and whether the child was biological or not was missing. This information was added starting with the data collection in 2011, changing the order of questions. Question 4 now covers the birth order, question 5 whether the child is the respondent's biological child or not. All other questions starting with former question 4, child's height and weight, are moved back by two positions.

bioage08 and bioage10

The ***bioage08*** and ***bioage10*** files are based on the parent-questionnaires 5 and 6, respectively. In the screening sample of 2010, question 13 was used in a slightly different version. This question covers the school conduct in both questionnaires and provides the information for the variable SCOLON1 through SCOLON9. Due to this change, valid observations for variables SCOLCON8 and SCOLCON9 exist for the screening sample in 2010 only, and hence these two variables are marked with a "SC 2010" in the variable label. An overview of items and respective differences is provided in table 9. As before, if a question is specific to one sample, observations stemming from the other sample will show the value "[-2] Does not apply" for this variable. Figures 1 and 2 show the questions as they appear in the respective questionnaires.

Table 9: *bioage08/bioage10* differences in SCOLON across samples

Variable Name	**Label Text**	**Item in Questionnaires 5/6**	
		All but Screening sample asked in 2010	**Screening sample, asked in 2010**
SCOLCON1	likes to go to school	1	2
SCOLCON2	doesn't get along with classmates	2	
SCOLCON3	thinks school is a waste of time	3	
SCOLCON4	never takes school work seriously	4	
SCOLCON5	is able to follow the lessons	5	3
SCOLCON6	doesn't get along with current teacher	6	
SCOLCON7	likes to study/has a zeal for learning	7	5
SCOLCON8	gets along with classmates (SC 2010)		1
SCOLCON9	gets along with current teacher (SC 2010)		4

Figure 1: Screenshot Parent Questionnaire 5 Screening Sample 2010
(Parent Questionnaire 6 similar)

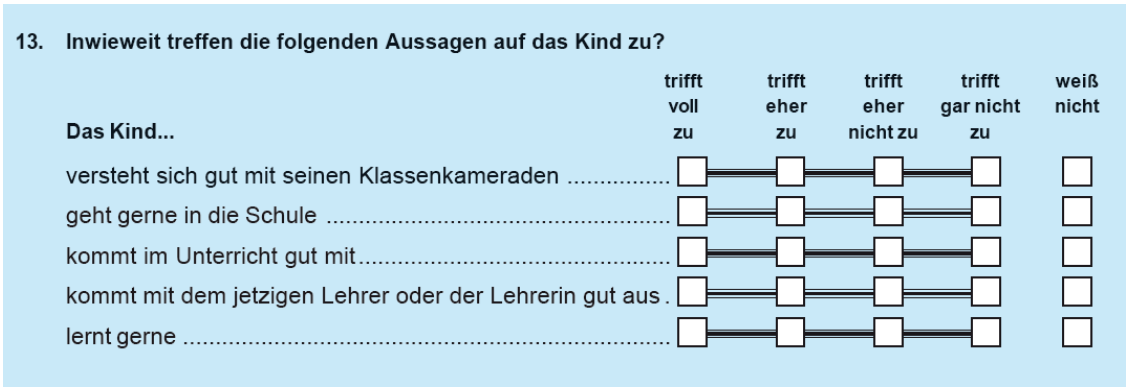
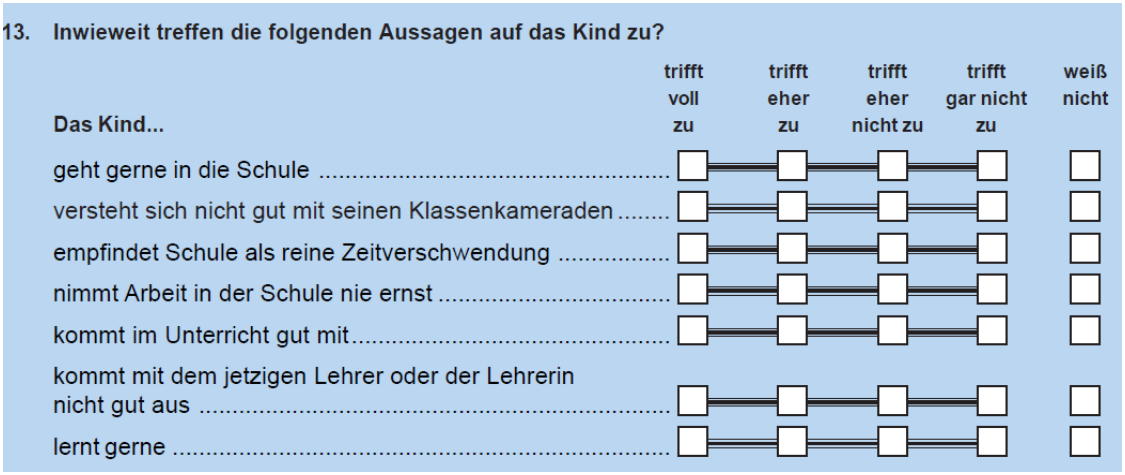


Figure 2: Screenshot Parent Questionnaire 5 Cohort Sample 2010 and 2011 sample
(Parent Questionnaire 6 similar)



bioage10

Until the wave 2012, the questions on the child's socio-economic behavior (Strengths-and-Difficulties, SDQ[11], captured in the variables B10BEHAV1-B10BEHAV25) have been asked on a 7-point scale ranging from 1 (does not apply at all) to 7 (fully applies). However, the international standard requires a 3-point scale, which will be implemented starting in 2013. In 2012, both scales have been used to find a way to easiliy "translate" the 7-point-scale into the 3-point-scale, mainly for the previous waves. To accomplish this, two questionnaire versions containing either scale were randomly distributed across households. The information from 2012 wave was then used to recode the information from 2010 and 2011 into a 3-point-scale in the ***bioage10*** datasets, while the original information remains available in the respective Parent-Questionnaires (***$eltern6***). Based on the distribution of each single item on the 3-point-scale, the 7-point-scale items can be categorized into the following 5 groups:

Table 10: Recoding scheme for SDQ questions in ***bioage10*** (BEHAV)

	Recoding scheme of item values	**# of items**
Group 1	"1" => "1"; "2"-"5" => "2"; "6"-"7" => "3"	4
Group 2	"1"-"3" => "1"; "4"-"5" => "2"; "6"-"7" => "3"	3
Group 3	"1"-"2" => "1"; "3"-"5" => "2"; "6"-"7" => "3"	10
Group 4	"1"-"2" => "1"; "3"-"6" => "2"; "7" => "3"	4
Group 5	"1"-"3" => "1"; "4"-"6" => "2"; "7" => "3"	3
Group 6	"1"-"1" => "1"; "3"-"4" => "2"; "5"-"7" => "3"	1

Following these schemes, all 25 variables were recoded. Before 2012, the questions on the child's socio-economic behavior only comprise 18 variables but were expanded to 25 variables in 2012. The "new" questions B10BEHAV3, B10BEHAV7-8, B10BEHAV11, B10BEHAV17-18 and B10BEHAV21 were not included in the ***bioage10*** datasets in 2010 and 2011 and have a missing code of -5 (Questions not included for sample) for the years 2010 and 2011 in the ***bioage10*** datasets from 2012 on. The following tables show the distributions in 2012 for a) the 7-point scales, b) the recoded 7-point-scales in 3-point-scale format, and c) the distribution for the 3-point-scale as found in the field.

[11] More information on the SDQ including translations from German to English can be found at http://www.sdqinfo.org/.

Table 11: Frequencies of original and recoded variables BEHAV1-BEHAV25

B10BEHAV1

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	1.02	1.02	1.88
2	2.04	41.84	44.77
3	4.59		
4	13.52		
5	21.68		
6	35.71	57.14	53.35
7	21.43		

B10BEHAV2

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	23.41	59.80	56.45
2	23.92		
3	12.47		
4	14.76	33.84	33.87
5	11.20		
6	7.89		
7	6.36	6.36	9.68

B10BEHAV3

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	41.58	68.88	67.02
2	27.30		
3	10.97	26.53	26.01
4	8.16		
5	7.40		
6	2.81	4.59	6.97
7	1.79		

B10BEHAV4

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	0.77	4.34	4.83
2	3.57		
3	6.12	42.86	40.21
4	14.54		
5	22.19		
6	29.85	52.81	54.96
7	22.96		

B10BEHAV5

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	25.26	46.17	51.61
2	20.92		
3	17.60	43.62	39.52
4	14.80		
5	11.22		
6	8.16	10.20	8.87
7	2.04		

B10BEHAV6

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	42.59	74.23	72.39
2	20.90		
3	9.20		
4	9.20	18.11	22.25
5	9.36		
6	5.77	7.65	5.36
7	2.96		

B10BEHAV7

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	1.79	6.38	5.11
2	4.59		
3	8.42	47.70	58.06
4	16.33		
5	22.96		
6	33.93	45.92	36.83
7	11.99		

B10BEHAV8

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	31.54	71.79	67.92
2	28.46		
3	11.79		
4	14.62	23.59	26.68
5	8.97		
6	3.33	4.62	5.39
7	1.28		

B10BEHAV9

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	2.29	2.29	2.96
2	3.31	24.68	22.37
3	3.31		
4	3.56		
5	14.50		
6	35.37	73.03	74.66
7	37.66		

B10BEHAV10

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	33.50	67.52	63.98
2	22.76		
3	11.25		
4	12.28	28.39	28.76
5	7.93		
6	8.18		
7	4.09	4.09	7.26

B10BEHAV11

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	3.05	3.05	1.88
2	1.53		
3	1.02	4.33	6.70
4	1.78		
5	3.31	92.62	91.42
6	13.99		
7	75.32		

B10BEHAV12

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	45.15	75.77	74.53
2	30.61		
3	8.42	23.98	24.66
4	8.93		
5	4.59		
6	2.04		
7	0.26	0.26	0.80

B10BEHAV13

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	44.78	74.05	78.17
2	29.26		
3	8.40	21.88	17.79
4	7.38		
5	6.11		
6	3.05	4.07	4.04
7	1.02		

B10BEHAV14

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	2.29	2.29	1.61
2	1.53	30.28	26.61
3	2.54		
4	10.69		
5	15.52		
6	38.42	67.43	71.77
7	29.01		

B10BEHAV15

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	14.50	35.37	35.66
2	20.87		
3	11.96	55.98	51.21
4	16.03		
5	15.01		
6	12.98		
7	8.65	8.65	13.14

B10BEHAV16

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	22.45	63.78	58.71
2	25.00		
3	16.33		
4	15.31	26.28	34.85
5	10.97		
6	7.91	9.95	6.43
7	2.04		

B10BEHAV17

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	2.30	4.86	3.23
2	2.56		
3	1.28	16.11	16.94
4	5.12		
5	9.72		
6	27.88	79.03	79.84
7	51.15		

B10BEHAV18

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	30.87	64.80	68.92
2	33.93		
3	12.50	33.93	28.65
4	11.48		
5	7.14		
6	2.81		
7	1.28	1.28	2.43

B10BEHAV19

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	36.69	65.63	70.70
2	28.94		
3	11.89	31.52	25.27
4	10.08		
5	6.20		
6	3.36		
7	2.84	2.84	4.03

B10BEHAV20

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	1.79	5.88	3.49
2	4.09		
3	5.88	37.34	48.26
4	13.81		
5	17.65		
6	30.95	56.78	51.74
7	25.83		

B10BEHAV21

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	75.51	90.82	90.32
2	12.24		
3	3.06		
4	3.57	7.91	9.14
5	3.32		
6	1.02		
7	1.28	1.28	0.54

B10BEHAV22

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	38.62	60.10	61.56
2	21.48		
3	9.97	36.32	34.14
4	19.95		
5	6.39		
6	2.81	3.58	4.30
7	0.77		

B10BEHAV23

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	32.48	60.87	64.34
2	28.39		
3	13.04	32.48	29.76
4	12.79		
5	6.65		
6	5.37	6.65	5.90
7	1.28		

B10BEHAV24

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	4.58	14.25	12.60
2	4.58		
3	10.18	49.36	52.01
4	25.70		
5	23.66		
6	21.63	36.39	35.39
7	9.67		

B10BEHAV25

	7-pt-scale	3-pt recoded	3-pt-scale orig
1	4.58	4.58	6.43
2	4.58	64.12	63.27
3	10.18		
4	25.70		
5	23.66		
6	21.63	31.30	30.29
7	9.67		

Documentation *biobirth*

Generated birth history for men and women

Stefan Damerow
(Based on the SOEP Documentation by Joachim Frick and Christian Schmitt)

Contents

General information

The data file ***biobirth*** provides a detailed overview of each individual's birth history, i.e. an account of biological children a woman has born or a man has fathered. While this information is gathered through the biographical questionnaire on the one hand, children are on the other hand also born during the panel life and then captured through the regular person interviews of mother and/or father. The parent questionnaires and partnership information from ***biocouply*** provide an additional set of information. ***biobirth*** combines these different sources into one consistent dataset. Thus, ***biobirth*** can be described as an accumulative data set, which contains the entire birth biography of all FiD respondents.

The ***biobirth*** file is based on every person, who had at least one successful FiD interview. It includes all individuals turning 17 during the survey year. ***biobirth*** covers the following information[12]:

(1) birth year and sex for each biological child up to the last date of interview[13]

(2) sum of children captured over the panel, split by their source – biography or other

(3) person identifier (KIDPNR) for each child, provided the child was ever living within a FiD household

(4) marital status and living circumstances (related to the biological mother/father) at the time of birth[14]

Information on data generation in biobirth

As mentioned above, there are various sources for the birth history. The variable KIDSOURCE[n] provides information on which source was used to generate the information for every of the respondent's children.

The main basis of the individual birth biography in ***biobirth*** is the information collected by the biography questionnaire. It provides information on the number of children, the birth date and the sex of each respondent's biological children. In addition, the relationship status at the time of each child's birth is recorded for every respondent. The biographical information is

[12] Up to FiD v2.0, the information where a child was living was included as well (variable KIDHOME). However, starting with distribution v2.1, this information was dropped, because it cannot be kept up to date. The information remains available in the ***$lela*** files. Whether a child still lives in the same household can be gathered using the yearly ***$pbrutto*** datasets.

[13] While the wave specific files ***$kind*** present the social and thus time-dependent parent-child relationships for children aged 16 or younger in the household, ***biobirth*** documents only biological parent-child relationships. This is the major difference to the SOEP version of this dataset, where the distinction between biological and adopted children is not possible.

[14] This information is not part of the SOEP distribution of ***biobirth***, as it is not asked in the questionnaire.

covered in the data set ***$lela***, filled out by each respondent once during his/her panel life. The birth history information is collected in the first part of the FiD ***$lela*** questionnaires ($LELTYP=1 or 3).

The ***$lela*** data sets do not contain the person number for children from the biographical information who are part of FiD (i.e. living in a household belonging to the FiD-sample). The dataset ***$kind*** provides the person number of the mother and father (MOTHNO$/FATHNO$) for every child in FiD. In combination with the indicator MOTHP$/FATHP$, biological child-parent relations can be identified to assign the child's person number (KIDPNR[n]) in ***biobirth***. Since ***$kind*** includes only children under 17 living in the household, the process generating these variables is extended to the whole data (see documentation of ***$kind*** for detailed information). Note that due to additional information with every new wave, the indicator MOTHP$/FATHP$ from previous waves may change in rare cases. In this case the birth biography is updated retrospectively.

The generating process of MOTHNO$/FATHNO$ and MOTHP$/FATHP$ is based not only on information from the birth biography. Parent – child relations are additionally found by the various parent questionnaires and in the variable $STELL (relationship to household head) from ***$pbrutto*** alone and in combination with partnership information from ***biocouply.*** As mentioned, the differences are stored in KIDSOURCE[n] for every child.

For children found by other sources except the birth biography, the information in KIDLITO[n] and KIDMAR[n] need to be obtained from other data. The dataset ***biocouply*** describes the partnership history of the respondent, so that it is possible to determine periods of marriage and coupled households. In case the child was born during a marriage and/or a partnership (same household), KIDLITO[n] and KIDMAR[n] are specified respectively.

The variable BIOKIDS sums up all respondent's children found by biography information while NEWKIDS covers the other sources.

The youth questionnaire is relevant for all young adults turning 17 during the survey year. This questionnaire is the first individual questionnaire a FiD-member answers for herself. In terms of ***biobirth***, it is important to note that the youth questionnaire – which itself does not contain any relevant information for ***biobirth*** – is answered *instead* of the biographical questionnaire. Assuming that only very few individuals give birth or father children before the age of 17, and that they can be identified within the household context (as long as they remain within FiD), this does not pose any problems for compiling the birth-biography of the respondents.

Once the children are found through the various ways, they are ordered chronologically, starting with the oldest child. In case of twins, the child with the lower person identifier is put first. If the twins are not part of the household, the order the respondent chose when reporting the children is maintained.

Finally, information on the birth history can be found through updates from the yearly questionnaire. Each individual is asked every year, whether a child “has come/born” into the household. With the help of the parent questionnaires and the previous observations, it is possible to identify new-born children in a household and relate them to their parents, and add information on the variables KIDLITO[n] and KIDMAR[n].

Variables in biobirth

BIOYEAR

Variable label	**"Year of biography survey"**
Value labels	BIOYEAR
Variable format	4-digit integer

Comment: In BIOYEAR, "-2 Does not apply" is given to individuals without the biographical information and for those who have answered the youth questionnaire instead of the biographical questionnaire. Accordingly the variables BIOAGE and BIOKIDS are set to "-2 Does not apply". Except for respondents with children from other sources, the same missing code is assigned to SUMKIDS and the children related variables (KIDPNR[n], KIDSEX[n], KIDYOB[n], KIDMOB[n], KIDMAR[n], KIDLITO[n], KIDSOURCE[n]).

BIOAGE

Variable label	**"Age when surveyed biography"**
Value labels	BIOAGE
Variable format	2-digit integer

Comment: For women/men who have not filled out the birth biographical questionnaire (yet), the code "-2 Does not apply" is used. Similarly individuals entering the dataset via the youth questionnaire are assigned this missing code.

BBSEX

Variable label	**"Gender of respondent"**
Value labels	SEX (1) male (2) female
Variable format	1-digit integer

SUMKIDS

Variable label	**"Total number of births"**
Variable format	1-digit integer

Comment: The variable is the total number of children identifiable within FiD by combining all available data up to the time of the last observation (SUMKIDS=BIOKIDS+NEWKIDS). If children are not mentioned by the respondent, or the source of information is not available (missing biographical information) the information in this variable may be underestimated. This is especially true for those respondents who did not fill in the biographical information.

BIOKIDS

Variable label **"Number of births from biography"**
Variable format 2-digit integer

Comment For women/men who have not filled out the birth biographical questionnaire (yet), the code "-2 Does not apply" is used. Similarly individuals entering the dataset via the youth questionnaire are assigned this missing code.

NEWKIDS

Variable label **"Number of births not from biography"**
Variable format 2-digit integer

Comment NEWKIDS provides the number of children identified through sources other than the birth biography. Hence all children born after the birth biography are stored here.

KIDSOURCE[n]

Variable label **"Source of birth biography"**

Value labels KIDSOURCE[n]
(1) birth biography
(2) parent questionnaire (E1 – E3 direct)
(3) parent questionnaire (E5/E6 direct)
(4) parent questionnaire (E1 – E3 indirect)
(5) biocouply
(6) $stell-variable ($pbrutto)
Variable format 1-digit integer

Comment The parent questionnaires contain direct and indirect information on the relationship to the child. On the one hand the respondent gives info about the own relationship (direct) to the child and furthermore his/her partner (indirect) when in household. The different sources in KIDSOURCE[n] can be ranked in relation to the quality of information. The best source is described by the direct info from parent questionnaire (E1 – E3, E5/E6) followed by the birth biography, then the indirect information, then biocouply and the information from the $STELL variable on the relationship to the household head from ***$pbrutto***.

KIDPNR[n]

Variable label **"Persnr [n]th child"**
Variable format 8-digit integer

Comment Provides the identifier for the first child [01] up to the fifteenth child *[15]*, if the child is included in the FiD-sample, i.e. living in one of the FiD-households. If not, i.e. the child has already left the household, is

living with a different parent, or has already died, and has never been part of FiD, this variable is set to "-2 Does not apply".

KIDSEX[n]

Variable label	**"Sex [n]th child"**
Value labels	KIDSEX[n] (1) male (2) female
Variable format	1-digit integer
Comment	n runs from 01 to 15.

KIDYOB[n]

Variable label	**"Year of birth child [n]"**
Variable format	4-digit integer
Comment	n runs from 01 to 15. If inconsistencies occur between ***$lela*** and ***$ppfad***, ***$ppfad*** is used as reference.

KIDMOB[n]

Variable label	**"Month of birth of child [n]"**
Variable format	2-digit integer
Comment	n runs from 01 to 15. If inconsistencies occur between ***$lela*** and ***$ppfad***, ***$ppfad*** is used as reference.

KIDMAR[n]

Variable label	**"Married at birth [n]th child"**
Value labels	KIDMAR[n] (1) Yes (2) No
Variable format	1-digit integer
Comment	n runs from 01 to 15.

KIDLITO[n]

Variable label	**"Lived with partner at birth [n]th child"**
Value labels	KIDLITO[n] (1) Yes (2) No
Variable format	1-digit integer
Comment	n runs from 01 to 15.

Note that for the variables KIDPNR[n], KIDSEX[n], KIDYOB[n], KIDMOB[n], KIDMAR[n] and KIDLITO[n] identical missing codes apply. The code "-2 Does not apply" is assigned, if there is no [n]th child found for this person. The code "-1 No answer" is used if information about the [n]th child is found but the specific information is missing.
For every woman/man a maximum of 15 entries for children is provided, although the biography questionnaire enables only ten possible entries regarding birth information. If there have been additional births up to the time the biography questionnaire is collected, they are recorded separately by the interviewer and are included in ***biobirth***. The sequence of children within ***biobirth*** is recorded with regards to the age of the children. The oldest child is recorded under KIDPNR01 the second oldest under KIDPNR02 and so forth.

Documentation *biomarsy* and *biocouply*

The Marital and Couple History Variables

Juliana Werneburg / Stefan Damerow

*This documentation is based on the comparable SOEP documentation **biomarsy** and **biomarsm** and has benefited from the work by Olaf Groh-Samberg and Florian R. Hertel.*[15] *For readability reasons, we do not specifically cite and specify text that has been used directly from the SOEP document.*

[15] Olaf Groh-Samberg and Florian R. Hertel (2010): The Marital History Files BIOMARSM and BIOMARSY. Chapter 5 in: Joachim R. Frick / Henning Lohmann (Eds.): Biography and Life History Data in the German Socio Economic Panel (SOEP, v26, 1984-2009), DIW Berlin.

Contents

General Information

"Familien in Deutschland" (FiD) provides individual marital and partnership histories in two data files, ***biomarsy*** and ***biocouply***. ***biomarsy*** is a spell dataset containing annual spells of the individuals' marital status, while ***biocouply*** provides annual spells on the partnership status. Both files comprise data on marital and couple biographies of respondents with a personal interview and additionally of adults living in an interviewed household without their own interview.

Both files contain whole relationship biographies starting at the year of birth. Thus, they mainly include retrospective information. In addition, they are extended by information given in every subsequent personal interview. The data file ***biomarsm***, which is known to SOEP users, is not provided in FiD, because there is only very little monthly information on relationship histories available yet. Additionally, the marital and couple status in the ***$pgen*** data files, stored as MARRST$$ and COUPST$$, are derived from ***biocouply*** for the time of the interview.

This documentation proceeds with a brief description of the two data files ***biocouply*** and ***biomarsy***. Users interested in more details may read further on how information on couple histories was collected in FiD and on the editing process of constructing logically consistent marital and couple histories.

Comparison with SOEP

FiD gathers more detailed information on a person's former relationships than the SOEP did up to 2011. For that reason, instead of only providing marital histories, FiD extended the data generation to whole couple histories provided in the dataset ***biocouply***. For the first time, the starting year of a relationship – independent of a later marriage – is available. In addition, a specific identification number for couples, COUPID, was created which will ease data handling for longitudinal analyses on relationships. Couple histories differentiate between being coupled vs. being married as well as between living together vs. living apart. Furthermore, separation periods for former couples who stay married afterwards can be analyzed. For convenience reasons the known dataset ***biomarsy*** is made available for FiD as well.

biocouply: A yearly couple biography

The spells in the data file ***biocouply*** contain retrospectively collected information on couple history since a respondent's year of birth on an annual basis. FiD's retrospective part of the person questionnaire covers up to four relationships in addition to the current status.

Information provided in the dataset is not only made available for respondents themselves. For most adults living in the same household, even if they were not interviewed, the current couple status can be reconstructed. In the case a respondent reported a relationship with a non-interviewed person living in the same household, information was copied to the non-responding partner. Persons aged 17 and interviewed only via Youth Questionnaire are not included in any of the datasets.

The data file contains thirteen variables: the household, couple and individual identifiers HHNR, COUPID and PERSNR as well as ten spell specific variables. The variable SPELLTYP documents the couple status with the possible categories ‘married, spouse in household’, ‘married, spouse not in household’, ‘coupled, partner in household’, ‘coupled, partner not in household’, ‘single‘, ‘separated’, ‘registered same-sex partnership, living together’, ‘registered same-sex partnership, living separately’. For same-sex partnerships it is only asked since wave 2 (2011), thus, this information cannot be reconstructed retrospectively. ‘Separated’ spells apply only to former couples who are still married, i.e. who are separated but not yet divorced. Note that ‘separated’ is a redundant spell in terms of completeness of couple histories: it always overlaps with other spells that contain the actual couple status(es) over the entire separation episode. Hence, by deletion of all separation spells the seamless couple history is preserved. The additionally assigned codes ‘unknown’ and ‘unit nonresponse’ indicate a lack of information for the respective period.

The variable SPELLNR is a chronological index number for each individual’s spell during the observation period. The variables BEGINY and ENDY provide the years in which a spell begins and ends, whereas the variables BEGIN and END indicate respondent’s age for users’ convenience. In ***biocouply***, spell systems for each individual always start with the respondent’s birth. Due to the fact that the spells’ duration is measured in years, it is important to note that an individual may encounter several events in the same year. In this case the variable SPELLNR allows the user to order spells with respect to the respondent’s life course. The SPELLTYP of the first spell per definition is ‘single’.

In addition, the indicator variables PDEATH and DIVORCE are provided. PDEATH indicates whether a respective spell ends with the death of a person’s partner. Single and gap spells are assigned a “(-2)” (does not apply). Please note that indicator PDEATH is not restricted to married persons, thus does not only refer to widowhood. By summing up PDEATH over the person’s life course, the number of subsequent states of widowhood can be retrieved easily (be careful to change missings into system missing values if you plan to sum up these indicators column-wise). DIVORCE works in a similar fashion. It indicates whether

the last marriage spell, that is the separated spell, ended in divorce. Hence, if it did not end in divorce it is coded as a still ongoing marriage. In this case ENDY is updated by year of last interview.

Variables in biocouply

The following provides an overview of the variables in ***biocouply***. Missing codes are not listed separately, as the usual conventions apply: "-1" refers to missing answers (don't know or refusal), "-2" indicates, that the question did not apply, whereas "-3" specifies implausible answers. There are some missing values (-1), (-2) or (-3) in BEGINY as well as in ENDY (BEGIN and END) indicating that we do not know the exact year of change in the couple status. Missing dates may simply indicate that the year was either not reported (-1), not asked for (-2) or implausible (-3), i.e. contradictory to other information. A gap might also have occurred: a) due to unit nonresponse or b) lack of substantial detail on an announced relationship or c) because it was not asked for. In order to differentiate the reasons for missing information the user can utilize variables REMARK and CENSOR.

The variable REMARK provides information on whether we had to edit or supplement original information provided by respondents in order to construct consistent couple biographies. Spells in ***biocouply*** are marked as 'edited' (in contrast to 'original') if the editing process involved substitution of or additions to original information as reported in the questionnaire. This happened, for example, if a respondent failed to report a relationship but valid information from a partner was available. Similarly, we inserted a divorce between two marriages, even if it was not specified, just because two marriages to separate persons at the same time are not legal (see later section for more details on editing). Furthermore, 'first spells' and 'gap spells' are marked separately. Note that inserted 'single' episodes between two consecutive reported relationships are edited as 'original spell'.

The variable CENSOR indicates whether a spell is left-censored, right-censored or censored on both sides. Furthermore, there is information included for the reasons of censoring. In principle, spells might be censored if they precede or follow a gap spell or if BEGIN or END is missing. The last spell for each person is marked as right-censored if a person is still in FiD and the current marital status is open ('last spell').

Table 1: Coding of variable CENSOR in ***biocouply***

Right: / Left:	not censored	censored missing	censored before gap	censored last spell	censored death
not censored	0	3	4	5	6

censored missing	1	7	8	9	10
censored after gap	2	11	12	13	14

COUPID

Variable label **"Couple identifier"**
Variable format 4-digit integer

SPELLNR

Variable label **"Consecutive Spell Number"**
Variable format 2-digit integer

Comment Range from (1) to (20)

SPELLTYP

Variable label **"Type of spell/event"**
Value labels SPELLTYP
(1) married, spouse in household
(2) married, spouse not in household
(3) coupled, partner in household
(4) coupled, partner not in household
(5) single
(6) separated
(7) registered same-sex partnership, living together
(8) registered same-sex partnership, living separately
(98) unknown
(99) unit nonresponse
Variable format 2-digit integer

BEGINY

Variable label **"Year spell begins"**
Variable format 4-digit integer

Comment Range from 1924 to 2013

ENDY

Variable label **"Year spell ends"**
Variable format 4-digit integer

Comment Range from 1939 to 2013

BEGIN

Variable label **"Age spell begins"**

Variable format 2-digit integer

Comment Range from 0 to 86

END

Variable label **"Age spell ends"**
Variable format 2-digit integer

Comment Range from 6 to 86

PDEATH

Variable label **"Death indicator"**
Value labels PDEATH
(0) spell does not end with death of partner
(1) spell ends with death of partner
Variable format 1-digit integer

DIVORCE

Variable label **"Divorce indicator"**
Value labels DICORCE
(0) spell does not end with divorce
(1) spell ends with divorce
Variable format 1-digit integer

CENSOR

Variable label **"Censoring information"**
Value labels CENSOR
(0) not censored
(1) LC no beginy
(2) LC after gap
(3) RC no endy
(4) RC before gap
(5) RC last spell
(6) RC death
(7) LC+RC no beginy a. no endy
(8) LC+RC no beginy a. before gap
(9) LC+RC no beginy a. last spell
(10) LC+RC no beginy a. death
(11) LC+RC after gap a. no endy
(12) LC+RC after gap a. before gap
(13) LC+RC after gap a. last spell
(14) LC+RC after gap a. death
(See Table above for detailed description on the censoring codes.)
Variable format 2-digit integer

REMARK

Variable label	**"Error Code"**
Value labels	REMARK (1) original spell (2) edited spell (3) gap spell (4) first spell
Variable format	1-digit integer

Sources of the couple history

For the construction of individual couple histories, information on the biography from the Biography Questionnaire ***$lela*** was mostly used. Figures 1 to 5 show those parts of the Biography and Person Questionnaire for the survey year 2011 which aimed at collecting respondents' couple history. In addition, the couple identifier COUPID and the generated partner pointer PARTP$$ of the generated dataset ***$pgen*** were used to link current partners living in the same household, to compare their answers, and supplement them if necessary.

Only long-term relationships, defined as lasting for at least six months, should be mentioned in the questionnaire (see Figure 1). For those interviewed again any changes that occurred during the last year were asked for (see Figure 2). From the Biography Questionnaire we obtain information on the current (see Figure 3) and up to three other marriages or relationships (see Figure 4) or three relationships plus one marriage (see Figure 5) that took place prior to the interview. Hence, all in all, up to five relationships are possible to record.

Figure 1: Introduction to marriage part of the Biography Questionnaire (translated)

The following part deals with relationships and marriage

All questions are related to relationships no matter whether you are married or not.

Whether you married in this relationship will be asked later.

These questions ask about both your current and your previous relationships.

We start with your current relationship. Afterwards, we ask retrospectively for the long-term relationship you had previously to the last-mentioned.

Concerning previous relationships, we call them long-term, if they lasted at least six months or longer.

When we ask about the previous long-term relationship, please always consider the relationship you had before the last mentioned, which lasted at least six months.

The amount of information collected about the different relationships – the current, the second to fourth previous relationship, and a former marriage if applicable – is not identical. Only in 2010, the year of moving together was not retrieved for the current relationship. For this reason, the year of moving together was estimated by the year of moving into the current household for current relationships. In those cases where both partners responded and gave different answers on the year of moving in, the most recent year was chosen as the date of

Figure 2: Questions on changes during last year (translated)

99. Has your family situation changed since December 31, 2010?
Please indicate if any of the following apply to you and if so, when this change occurred.

	Yes	**2011** in month	**2010** in month
Started a new relationship.....	☐	…☐☐	…☐☐
Got married…….....	☐	…☐☐	…☐☐
Moved in with my partner……....	☐	…☐☐	…☐☐
I separated from my spouse / partner ...	☐	…☐☐	…☐☐
I got divorced.....	☐	…☐☐	…☐☐
My spouse / partner died……...	☐	…☐☐	…☐☐

Figure 3: Questions on the current relationship (translated, Person Questionnaire 2011)

L43. What about the present: Do you currently have a long-term relationship?

Yes...☐ (→ L44) No…☐ (→ Go to question L53)

L44. When did the relationship with this partner start?

Year ☐☐☐☐

L47. Does your partner live in this household?

Yes...☐ (→ L48) No…☐ (→ Go to question L49)

L47a. When did you move in with your partner? *[Asked in 2011 only]*

Year ☐☐☐☐

L48. Please tell the first name of your partner.

_____

L49. Did you live together with that partner in the past?

Yes...☐ (→ L50) No…☐ (→ Go to question L45)

L50. When did you give up your common accommodation or when did you or your partner move out?

Year ☐☐☐☐

L45. Are your married to this partner?

Yes...☐ (→ L46) No…☐ (→ Go to question L51)

L46. When did you marry?

Year ☐☐☐☐

Figure 4: Questions on previous relationships (translated)

Now we cover your previous relationship – that is the one prior to the current / the last mentioned relationship.
(Analogously: 2nd to 4th relationship)

L53. Did you have a previous long-term relationship?

Yes...☐ (→ L54) No…☐ (→ Go to question L63)

L54. When did that relationship start?

Year ☐☐☐☐

L55. When and in which way did that relationship end?

Year ☐☐☐☐

Through separation……☐ (→ L56)

Through death……☐

L56. Did you live with this partner?

Yes...☐ (→ L57) No…☐ (→ Go to question L59)

L57. When did you move in with this partner?

Year ☐☐☐☐

L58. And when did you give up your common accommodation or when did you or your partner move out?

Year ☐☐☐☐

The common domicile was not vacated……☐

L59. Did you marry this partner?

Yes...☐ (→ L60) No…☐ (→ Go to question L63)

L60. When did you marry?

Year ☐☐☐☐

L61. Did you get a divorce?

Yes...☐ (→ L62) No…☐ (→ Go to question L63)

L62. When were you divorced?

Year ☐☐☐☐

Figure 5: Questions on other marriage (translated)

(only if respondent was not married in the previous three / four relationships)

L62a. Have you ever been married in another relationship that was not yet mentioned?

Yes...☐ (→ L62b) No…☐ (→ Go to question L63)

L62b. When did you marry in this relationship?

Year ☐☐☐☐

L62c. Is this marriage still persisting?

Yes...☐ (→ L62d) No…☐ (→ Go to question L63)

L62d. When and in which way did that relationship end?

Year ☐☐☐☐

Through separation……☐

Through death……☐

moving in together. Information on a previous marriage is just elementary (see also Figure 5). Moreover, it was asked for only in the case that three relationships had already been mentioned and none of these were marriages.

Construction of couple history

Information on couple history mainly stems from respondents' retrospective reports on their own history. Thus, no other benchmark on the substance of these reports exists. This led to inconsistencies – for example, overlaps of two relationships occurred, or the ordering of becoming a couple, moving together and marrying was reported differently than expected. Even though many different patterns of the reported histories occurred, sometimes with unusual appearances, they are still possible for the most part. Hence, no verified decision between measurement error and uncommon reality can be made. For that reason, whenever possible, couple histories are left as they were reported. Only in rare cases, restrictions, corrections on orderings of events or changes of reported years are conducted (read the following carefully for details). Little original information was changed (note: to retrieve original data, see the data file originating from the Biography Questionnaire, ***$lela***). Further corrections to smooth out irregularities are thus left to the user. Note that consistency checks between waves are done as well. That way, changes between data distributions are possible for former waves.

A clear ordering of spells needs to be sustained if information on timing is missing. Besides, information was asked on an annual basis. Thus, ordering of those spells cannot be done empirically, but has to be decided in advance. For that reason the following default rules were obeyed to obtain logically consistent histories if no other information forced to do otherwise:

1. Every individual couple history starts with the state 'single'. In general it is assumed to last at least until the age of 15. However, we did not restrict age within a relationship, that is, unmarried relationships are allowed to start anytime between the age of 0 and 15. We restricted the age of marriage to be at least 15, though.
2. Every spell set for a certain couple starts with the state 'coupled, partner not in household'. One exception exists: if respondents report a year of moving together that lies before the start of their relationship, this specific couple history starts with 'coupled, partner in household'. Note that in this case the information when this couple moved together is not available in ***biocouply*** anymore. You would have to look it up in the original source stored in the data file ***$lela***.
3. If there is no evidence to the contrary, it is assumed that married couples live together and moved together before marriage. That is: for married couples their specific couple history starts with 'coupled, partner not in household' and is followed by a spell 'coupled, partner in household' before their marriage spell 'married, spouse in household' starts. Thus, if a couple moved together in the same year they married, a spell 'coupled, partner in household' is included anyway. Additionally, this assumption applies to a marriage reported as the fourth/fifth reported relationship (see Figure 4). Note that dates of becoming a couple and moving together or whether they moved together at all are not known in that case.
4. To ensure that any information possible is included into the spell dataset, a spell 'married, spouse not in household' is created even if the date of moving out and the end of a relationship fall into the same year or if either of the two dates is not known. Likewise, if a partner moves out of the joint household in the year of marriage, the spell 'married, spouse in household' is included anyway, even though it may be redundant. Note that the date of moving out is lost if it was reported to come after the end of a relationship. (It can still be retrieved from the original source stored in the data file ***$lela***.)
5. Any (formerly) married couple that is not an active relationship anymore, i.e. married who are separated but not yet divorced, ends with a spell 'separated'. As long as they are not divorced yet, the end date of those separated spells is the same as their last

interview year. Again, one exception exists: the separation spell is not added if marriage clearly ended with the death of the respondent's partner. If not known, a 'separated' spell is added. Note that 'separated' is a redundant spell in terms of couple history: it always and fully overlaps with other spells that contain the actual couple status(es) over the entire separation episode. Hence, by deletion of all separation spells, the seamless couple history is preserved. It should be noted that information on how a marriage ended – via death or divorce – is saved in the separation spells in PDEATH and DIVORCE, hence care should be applied when deleting those spells.

6. Because ***biocouply*** documents couple statuses and not marital statuses, it is possible to become single after marriage (in ***biomarsy*** the only possible change from 'divorced' or 'widowed' is to 'married').
7. As it was possible to mention another relationship in the questionnaire (up to four were allowed, see above), it is assumed that periods between those relationships mentioned were 'single' states and thus filled accordingly. Note that this relates to long-term relationships only.
8. Information on the death of a partner or a divorce is stored in variables PDEATH and DIVORCE. If not applicable, that is if a person is currently single, PDEATH and DIVORCE are set to "-2", does not apply. If a respective couple is not married, DIVORCE is set to "-2", does not apply, as well. (Be careful to change missings into system missing values if you plan to sum up these indicators column-wise.)

For many non-interviewed persons information on the current relationship can be reconstructed, because the respective partner is available from his or her personal interview. Thus, those non-interviewed adults remain in the ***biocouply*** dataset, even though information is not directly retrieved from them. If information on their current couple status was given by their partner, it is fully copied to these non-respondents as well. Concerning their prior history, the previously mentioned rules are followed: from age 0 to 15 they are stated 'single'. From age 15 until the current relationship a gap – a spell with SPELLTYP 'unit nonresponse' (99) – is included. Non-interviewed persons, for whom no information is available get a gap – a spell with SPELLTYP 'unit nonresponse' (99) – from age 15 onwards.

Consistency checks are possible in several ways: a) contradictions between information given by a respondent and her relationship to the head of household (see $PSTELL in the dataset ***$pbrutto***). This information was given by the head of household in the Household

Questionnaire and processed in the partner pointer PARTP$$. b) for the current relationship answers of both partners can be compared if a possible partner was identified via COUPID and both were interviewed; c) dates within a personal couple history might appear plausible or not. a) and b) are accomplished and single-case corrections done if reasonable. Concerning c) it is tried to leave as much of the original information as it is given by the respondent to allow for analyses of uncommon patterns as well. Some overlaps and inconsistencies might still appear in the dataset, which you may want to remove. Generally, inconsistencies were dealt with the following way:

1. If a respondent reported an early marriage in the Biography Questionnaire, we restricted its beginning to age 15. But we did not restrict age within a relationship, that is, unmarried relationships are allowed to start anytime between the age of 0 and 15.
2. Within a persons' reported history, dates are not edited or overwritten (such as dates of marriage or begin of relationships). Thus, it is possible that a person is married more than once at the same time or has multiple, overlapping relationships in a certain period. That is, relationships do not need to be consecutive. The variable SPELLNR provides the sorting order, whereby the most recent spell comes last, that is, the relationship that ended most recently (no matter which are the starting dates). If no clear dates were available, the sorting in SPELLNR reflects the order of reporting by the respondent.
3. Within a specific couples' reported history, dates are not edited or overwritten. E.g. it was not corrected if the start of a marriage was reported to be before the start of a relationship. This way 7 spells occur with a starting date coming after the end date of the spell. As mentioned above, it was accepted if the date of moving together was before the year of becoming a couple. In this case, this specific couple history starts with 'coupled, partner in household', while the information when this couple moved together is not available in ***biocouply*** anymore (it is available via the data file ***$lela***).
4. Contradictions in both partners' answers are left as they were, even if both were interviewed in the current wave. E.g. information on starting date or the date of marriage of their joint relationship often differs. As before, neither date is edited or overwritten between both partners of a current couple. To check for contradictions you can link them via couple identifier COUPID.
5. For very few cases, there were contradictions about the actual partnership between to individuals in the sample. E.g., if one part of a couple or the position to the head of household suggested a couple in a household, but one or both answered to be single or

coupled with someone else (outside the household), it was tried to link them by COUPID in their respective former relationship. Note that those single case corrections are not flagged if they just concern the couple identifier COUPID.

Some respondents refused to answer some or all questions in the relationship part. To some extent, missing values and contradiction are similar in the generating sense. Hence, some of the following rules applied to missing values may seem familiar:

1. In some cases it is possible that a respondent might have had earlier relationships, but they could not be named anymore due to the above mentioned questionnaire restriction. Thus, it is not known whether there were other partners before the last mentioned. In these cases, gaps were introduced, that is a spell 'unknown' (98) was inserted.
2. If a respondent stated to be married in a relationship, but it is not known whether the couple lived together as well, the above mentioned default course is assumed. In some rare cases details on the relationship are not known, as just a relationship was reported. There thus exists no information whether the couple moved together or whether they were married. Here, gaps, (98) 'unknown', are introduced as well.
3. For current relationships with both partners living together, it is not known when they moved together. In these cases the year of moving together was substituted by the year of moving into the current household. If both partners were respondents and gave a differing answer on the year of moving in, the most recent year was chosen as the date of moving together. If the date of moving into the household comes before the year the relationship started, it is assumed that both partners moved together the very same year they became a couple. The user may decide herself whether a very long time span between becoming a couple and the estimated year of moving together might indicate that the couple already changed households together via moving.
4. Some interviewed persons did not report a current but – given other information – likely relationship. This contradiction is apparent either by the information of the related partner, by the relation to the head of household or seen in case-to-case checks, e.g. via common children identified through the parent questionnaires. For these cases, the relationship is included into the person's relationship record with a default record fitting to the identified marital status, but all dates are set to (-1), 'no answer'. The decision to copy the dates from the corresponding reporting partner is left to the user.

biomarsy: A yearly marital biography

The data file ***biomarsy*** provides retrospectively collected information on marital history starting with the respondent's year of birth. ***biomarsy*** in FiD is fully comparable to the version given in the SOEP: data are presented in spell format on an annual basis. ***biomarsy*** is derived by collapsing the dataset ***biocouply*** while applying only minor changes to the format. Thus, both datasets are consistent with each other, where ***biomarsy*** is a subset of ***biocouply***, focusing only on marriage, widowhood and divorce. Like ***biocouply***, ***biomarsy*** includes both respondents and non-responding adults living in an interviewed household.

The ***biomarsy*** file comprises eleven variables, including the individual and household identifiers HHNR and PERSNR as well as SPELLTYP. The variable SPELLTYP documents the marital status as known from the SOEP. It has the possible categories 'not married', 'married', 'divorced', widowed or divorced' and 'gap'. Once married, a later spell 'not married' is not assigned anymore. Note that we renamed the known SOEP code "1" 'single' to 'not married'. This is to indicate that it is possible the respondent might have a partner anyway. If you are interested in this information, we recommend using ***biocouply*** instead of ***biomarsy***.

SPELLTYP has one category 'divorced or widowed' in ***biomarsy***, which indicates that a marriage definitely ended, although we do not know whether this happened due to divorce or due to the death of the spouse, as this information is missing in the Biography Questionnaire. Because spell durations are measured in years, it is important to note that an individual may encounter several events in the same year. In this case the variable SPELLNR allows the user to order the spells with respect to the respondent's life course. The variables BEGINY and ENDY provide the years in which a spell begins and ends, while the variables BEGIN and END indicate the respective age of the respondent for users' convenience. The spell system for each individual in ***biomarsy*** always starts with the birth of the respondent. We thus created a first spell for each individual ever interviewed in the SOEP starting in the year of birth and continuing at least until the year in which a person reaches the age of 15. The SPELLTYP of the first spell per definition is 'not married'. Even if a respondent reported an earlier marriage in the Biography Questionnaire, we restricted its beginning to age 15.

Variables REMARK and CENSOR are constructed as in ***biocouply***. Please see the explanations above.

The sources of ***biomarsy*** are identical to those in ***biocouply***. Please consult the documentation of ***biocouply*** for further details.

Variables in biomarsy

SPELLNR

Variable label	**"Consecutive Spell Number"**
Variable format	1-digit integer
Comment	Range from (1) to (8)

SPELLTYP

Variable label	**"Marital status"**
Value labels	SPELLTYP (1) not married (2) married (3) divorced (4) widowed (5) divorced or widowed (6) registered same-sex partnership (9) gap
Variable format	1-digit integer

BEGINY

Variable label	**"Year spell begins"**
Variable format	4-digit integer
Comment	Range from 1924 to 2013

ENDY

Variable label	**"Year spell ends"**
Variable format	4-digit integer

BEGIN

Variable label	**"Age spell begins"**
Variable format	2-digit integer
Comment	Range from 0 to 86

END

Variable label	**"Age spell ends"**
Variable format	2-digit integer
Comment	Range from 6 to 86

CENSOR

Variable label	**"Censoring information"**
Value labels	CENSOR

(0) not censored
(15) LC no beginy
(16) LC after gap
(17) RC no endy
(18) RC before gap
(19) RC last spell
(20) RC death
(21) LC+RC no beginy a. no endy
(22) LC+RC no beginy a. before gap
(23) LC+RC no beginy a. last spell
(24) LC+RC no beginy a. death
(25) LC+RC after gap a. no endy
(26) LC+RC after gap a. before gap
(27) LC+RC after gap a. last spell
(28) LC+RC after gap a. death
(See Table above for detailed description on the censoring codes.)

Variable format 2-digit integer

REMARK

Variable label "**Error Code"**
Value labels REMARK
(1) original spell
(2) edited spell
(3) gap spell
(4) first spell
Variable format 1-digit integer

Construction of marital histories

Marital histories are derived directly from ***biocouply***. Hence, the information used is identical to ***biocouply***, although shorter and sometimes in a different format. The following rules hold:

1. Every individual marital history starts with the state 'not married'. We did not allow a person to be married before age 15.
2. From 'not married', one can only change to 'married'.
3. There is no possible return to 'not married' once a person was ever 'married'. The only possible change from 'married' is to 'divorced' or 'widowed'.
4. The only possible change from 'divorced' or 'widowed' is to 'married'.

Documentation *$kind*

Person-related variables on children (up to the age of 16)
within the household

By Stefan Damerow and Moritz Mannschreck
(Based on the SOEP Documentation by Joachim Frick)

General information

The variables from the annual ***$kind*** files described in the following are not based on answers provided by the children themselves but by answers to the household questionnaire provided by the respondent within the household (head of household). This data are disaggregated on the person level and saved as child-specific entries in the file ***$kind***.

The annual ***$kind*** datasets also contain additional information on institutional care, school attendance and extracurricular activities for children up to age 16. Questions on the costs of childcare as well as non-institutional care arrangements are also asked. These variables are not described here. The variable names correspond to the numeration in the household questionnaire.

The persons in ***$kind*** are identifiable with the Codes "120 to 126" in the wave-specific variable $NETTO (in the file ***ppfad***).

List of variables:

Appendix

PSAMPLE

Variable label	**"Subsample"**
Value label	PSAMPLE (61) FiD 2007 Birth Cohort (62) FiD 2008 Birth Cohort (63) FiD 2009 Birth Cohort (64) FiD 2010 Birth Cohort (65) FiD Screening (sampled 2010) (66) FiD Screening (sampled 2011)
Variable format	2-digit integer
Comment	Note that this variable is included in all datasets, and provides information whether the household or person originates from the cohort or the screening sample in FiD. In ***ppfad*** it is also named PSAMPLE, in ***hpfad*** it is named HSAMPLE; in all other datasets it is called SAMPLE1, which is analogous to the SOEP notation.

$WELLE

Variable label	**"Wave"**
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	This variable reports the survey year.

$KGJAHR

Variable label	**"Year of birth"**
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Year of birth of the child (4-digit). The value of $KGJAHR can in some cases differ from the longitudinally tested data in the central file ***ppfad*** (variable GEBJAHR).

$KGMON

Variable label	**"Month of birth"**
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	The child's month of birth. The value of $KGMON can in some cases differ from the longitudinally tested data in the central file ***ppfad*** (variable GEBMONAT).

$KSEX

Variable label	**"Sex"**
Value label	$KSEX (1) male

(2) female

Variable format 1-digit integer

$ - Wave F10, F11, F12, F13

Comment Child's sex. The value of $KSEX can in some cases differ from the longitudinally tested data in the central file ***ppfad*** (variable SEX).

$KSTELL

Variable label **"Relationship to head of household"**

Value label $KSTELL

Waves F10 and F11

(0) Head Of Household (HH)
(1) Spouse of HH head
(2) Partner of HH head
(3) Son, daughter of HH head
(4) Foster child of HH head
(5) Son, daughter-in-law of HH head
(6) Father, mother of HH head
(7) Parent-in-law of HH head
(8) Sister, brother of HH head
(9) Grandchild of HH head
(10) Other relative of HH head
(11) Not related to HH head
(12) Stepchild of HH head
(13) Same-sex spouse
(99) Unknown

Wave F12 and F13

(0) Household (HH) head
(11) Spouse of HH head
(12) Same-sex spouse
(13) Life partner of HH head
(21) Son, daughter of HH head
(22) Stepchild of HH head
(23) Adopted child of HH head
(24) Foster child of HH head
(25) Grandchild of HH head
(26) Great-grandchild of HH head
(27) Son, daughter-in-law of HH head
(31) Father, other of HH head
(32) Stepmother, stepfather of HH head
(33) Adoptive mother, father of HH head
(34) Foster mother, father of HH head
(35) Parent-in-law of HH head
(36) Grandmother, gradfather of HH head
(41) Sister, brother of HH head
(42) Half-sister, -brother of HH head
(43) Stepsister, -brother
(44) Adopted sister, brother of HH head

(45) Foster sister, brother of HH head
(51) Sister-in-law/brother-in-law 1: spouse or life partner of HH head's sibling
(52) Sister-in-law/brother-in-law 2: sibling of HH head's spouse or life partner
(61) Aunt/uncle of HH head
(62) Niece/nephew of HH head
(63) Cousin of HH head
(64) Other relative of HH head
(71) Not related to HH head
(99) Unknown

Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Copy of the data from $STELL from the wave-specific file ***$pbrutto***. Starting in F12, $STELL was changed and includes more values than in previous waves, which allows more precise identifications of relationships within the households. The values (0)-(2) as well as (6)-(7) in 2010 and 2011 and (0)-(13) and (31)-(36) in 2012 and 2013 are by definition not a part of the file ***$kind***.

$KINHH

Variable label	**"Household Membership"**
Value label	$KINHH
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Copy of the data from $PZUG from the wave-specific file ***$pbrutto***. In ***$kind*** some of the value labels do not apply due to the age of child.

$HHGR

Variable label	"**Number of persons in HH"**
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Copy of the variables $HHGR from the wave-specific files ***$hbrutto***.

$KZAHL

Variable label	**"Number of children in the HH"**
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	For each child in ***$kind***, the variable $KZAHL provides the total number of children up to age 16 in the current household

MOTHNO$

Variable label	**"Never Changing Person ID (mother)"**
Variable format	8-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

Person ID number of the child's mother. MOTHNO$ is comparable to the SOEP Variable $KMUTTI, except that in addition to biological and adoptive mothers, it includes social and foster mothers (for definitions see MOTHP$).
The identification of mother-child relation starts with the wave-specific files ***$lela***, in which the birth biography includes information about birth date, sex and the first name (not in data distribution) of the respondent's children. Since this information is unique for all members of a household, it is possible to identify children in the wave-specific files ***$pbrutto*** and assign the child's PERSNR. Within that procedure, problems occur if specific information is missing or obviously wrong in the biographical data. With the help of $STELL (= relationship to HH head in ***$pbrutto***, see appendix) potential children can be found. If a child could not be matched directly but is the "right" child, based on the name, birth date and sex, the person ID is matched. Furthermore, $STELL is used as the second source for the mother-child relation as ***$lela*** may not be available for every potential mother (see appendix). To obtain the person ID of social mothers, the biological father's partner is derived from the variable PARTNR$$ (=person ID number of spouse or partner), which is taken from the wave-specific files ***$pgen*** and the $STELL variable. Furthermore, the variable $STELL identifies foster mothers.

MOTHP$

Variable label	**"Indicator for mother's relationship to the child"**
Value label	MOTHP$ (1) biological (2) social (3) adoptive (4) foster (5) social, same-sex partners (6) unknown
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13

Comment

Indicator for the mother-child relation in the household. As described in MOTHNO$ the mother-child relation are linked with the person ID of the mother and the child. Within the procedures of identification the indicators (for mothers and fathers) are defined according to the source as follows:
If data from the wave-specific file ***$lela*** is available, first, it is possible to differentiate between "biological" and "adoptive" (Variable $l064g* in ***$lela***). Secondly, the mother can be defined as the social mother of her partner's biological children when she has not mentioned them in ***$lela***, but lives in the same household.

The identification by $STELL in the wave-specific file ***$pbrutto*** can only be used to identify foster mothers, because it is not possible to make a difference between biological, adoptive and step-children in $STELL (see appendix). Thus, potential biological relations for whom information is only found in $STELL are defined as "unknown". For further identification of biological mothers the individual parent questionnaires are used as they include information about motherhood (as generated variable BIOCHILD in files ***bioage01***, ***bioage02***, ***bioage03***, and ***bioage06*** and variable BIOPAR in file ***bioage08*** and ***bioage10***).
Finally the dataset ***biocouply*** gives information about the partnerships, specifically their beginnings. Unknown mother-child relations found by $STELL are replaced by the indicator "biological" when the child's date of birth falls in the history of the current partnership.
Due to additional information with every new wave, the indicator from previous waves may change in rare cases. In this case the MOTHP$ is updated retrospectively.
Note that there are a few cases of same-sex partnerships in FiD. In the event that these are women, MOTHNO$ is set to the PERSNR of the woman with the highest value in MOTHP$, while FATHNO$ is taken from the other partner (i.e. the biological mother would be found in MOTHNO$, the social mother in FATHNO$). If it is a male couple, FATHNO$ is the PERSNR of the man with the highest value in FATHP$, while MOTHNO$ is taken from the other partner. As these cases are rare, we refrained from adding a sex indicator for MOTHP$ and FATHP$. This information can always be obtained through ***ppfad***.

FATHNO$

Variable label	**"Never Changing Person ID (father)"**
Variable format	8-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	Variable generated identically to MOTHNO$, see description above.

FATHP$

Variable label	**"Indicator for father's relationship to the child"**
Value labels	FATHP$ (1) biological (2) social (3) adoptive (4) foster (5) social, same-sex partners (6) unknown
Variable format	1-digit integer
$$ - Survey Years	$$=10, 11, 12, 13
Comment	Indicator for the relationship father-child in the household. Variable generated according to MOTHP$, see description above.

schltypE$

Variable label **"Schooltype (generated)"**
Value labels SCHLTYPE$
(1) primary school
(2) secondary general school
(3) intermediate school
(4) upper secondary
(5) comprehensive school
(6) vocational school
(7) school with special pedagogical concepts
(8) special needs school
(9) other
Variable format 1-digit integer
$$ - Survey Years $$=10, 11, 12, 13

Comment Question 57a of the household questionnaire (survey years 2010/2011) and question 62a (survey years 2012/2013) specify the school the respective child goes to, with the option of an open answer in "other". As many respondents use this field to put in specific schools like Waldorf, Montessori or schools for children with special needs, this additional information is recoded in this variable. Note that the recoded cases all have "-2 Does not apply" in the original variable.

Appendix

Conversion of $STELL from 2012/2013 to 2010/2011 Version

$STELL in waves F10 and F11		**$STELL in wave F12**	
0	Head Of Household (HH)	0	Household (HH) head
1	Spouse of HH head	11	Spouse of HH head
2	Partner of HH head	13	Life partner of HH head
3	Son, daughter of HH head	21	Son, daughter of HH head
		23	Adopted child of HH head
4	Foster child of HH head	24	Foster child of HH head
5	Son, daughter-in-law of HH head	27	Son, daughter-in-law of HH head
6	Father, mother of HH head	31	Father, other of HH head
		33	Adoptive mother, father of HH head
		34	Foster mother, father of HH head
7	Parent-in-law of HH head	35	Parent-in-law of HH head
8	Sister, brother of HH head	41	Sister, brother of HH head
		42	Half-sister, -brother of HH head
		43	Stepsister, -brother
		44	Adopted sister, brother of HH head
		45	Foster sister, brother of HH head
		51	Sister-in-law/brother-in-law 1: spouse or life partner of HH head's sibling

		52	Sister-in-law/brother-in-law 2: sibling of HH head's spouse or life partner
9	Grandchild of HH head	25	Grandchild of HH head
10	Other relative of HH head	26	Great-grandchild of HH head
		32	Stepmother, stepfather of HH head
		36	Grandmother, gradfather of HH head
		61	Aunt/uncle of HH head
		62	Niece/nephew of HH head
		63	Cousin of HH head
		64	Other relative of HH head
11	Not related to HH head	71	Not related to HH head
12	Stepchild of HH head	22	Stepchild of HH head
13	Same-sex spouse	12	Same-sex spouse
99	Unknown	99	Unknown

$STELL-Combinations to identify mother-child relationships

Potential mother-child relationships as a combination of the variable $STELL in waves F10 and F11 (accordingly for father-child relationships, conversion table above can be used to translate mother-child relationships to waves F12 and F13)

$STELL of the		**Potential mother-child relationship**
woman	**another person**	**(reference person = head of the HH)** **In this case the person is the…**
0 1	3 3	Child of reference person Child of the wife of reference person
1 1	11 12	Child of the wife of reference person, but not child of reference person
2 2 2	3 11 12	Child of "life companion" of reference person and of reference person Child of "life companion" of reference person but not of reference person
3 4 5	9 9 9	Child of daughter of reference person Child of foster child of reference person Child of daughter in law of reference person (3 generation household)
6 6	0 8	Child is reference person, lives with his mother in the same household Child is the sister / brother of reference person, the siblings live with their mother in the same household
7 7	1 8	Child is spouse of reference person and lives together with spouse and mother in the same household Child is daughter / son of the mother in law of reference person, but not the spouse of the reference person rather the sister in law / brother in law of reference person
8 9 10	10 10 10	Child is niece / nephew of reference person, mother is sister / sister in law of reference person Child is another relation to reference person, great grandchild of reference person Mother and child have another relation to reference person
11	11	Child and mother are in no way related to reference person

Documentation *biojob*

Detailed Information on First and Last Job

Moritz Mannschreck

*This documentation is based on the comparable SOEP documentation on **biojob** and has benefited from previous work of Tanja Schmidt, Hansjoerg Haas, Anita Kottwitz, Daniel Wachtlin, Mathis Schroeder and Thorsten Schneider. For readability reasons we do not specifically cite and specify text that has been used directly from the SOEP document.*

General information

biojob provides detailed information on the respondent's first and last job. The relevant information is taken from the $***lela***-files which contain biographical information for all FiD respondents who are 18 or older in their first wave. ***biojob*** consists of generated variables as well as plain questionnaire information. Concerning different sources of information, the following priority scheme is applied: First the plain information stemming directly from questions on the relevant topic in the latest valid $***lela***-file is used. In case of inconsistencies, which will be explained later on, the latest valid information stemming from the ***pbiospe***-file is also used. The ***pbiospe***-file consists of spell data concerning the retrospective question 'what did you do since the age of 15' in the Biography Questionnaire as well as the question on activities in the last year in the Individual Questionnaire (for detailed information see ***pbiospe*** documentation). In contrast to the SOEP, ***biojob*** does not contain information from the Youth Questionnaire which is answered by respondents aged 16 to 17. Due to missing information, ***biojob*** contains fewer variables than the SOEP-version of ***biojob***.

For all variables, the information provided looks as follows:

Variable label	Provides the label of the variable as it is given in the dataset. Variables are given in CAPTIAL letters, even though they might appear in small letters in the dataset. This is simply for readability.
Value labels	*LBLNME* In case *VARNME* is categorical, *LBLNME* specifies the labels for each category, and the value labels are listed here. Note that the standard missing value labels (-1: No answer; -2: Does not apply; -3: Not valid) are not listed, but apply to all variables in this dataset.
Variable format	Specifies the format for each variable, e.g. "1-digit integer" or "string".
Comment:	Provides more detailed information on the generating process, also on the population the variable is specified for, if necessary. Here, variables used, changes between waves, or any other anomalies are mentioned and their relevance explained.

If you have questions regarding ***biojob*** data for the FiD-distribution, unless noted otherwise, please contact Mathis Schröder at +49 (0)30 / 89789 - 222.

List of variables:

AGEFJOB

Variable label **"Age at first job"**
Variable format 2-digit integer

Comment The variable AGEFJOB provides the age at entry into the working force.

In the Biography Questionnaire people either have to give information on their age at entry into the working force or have to state that they have never worked before the time of the interview. The latter information is used in the variable NOJOB. Some respondents have low values on AGEFJOB. Since most of the respective respondents worked either in low skilled blue collar or white collar jobs or within the family business, we consider the information as valid.

AGEINFO

Variable label **"Info Source Age First Job"**
Value labels AGEINFO
(1) LELA-Files (case (a) below)
(2) PBIOSPE Age First Job LT 15 BEGIN GT 15 (c)
(3) PB Not Worked And Later Begin (b)
(4) PB N.W. And Earlier Begin (d)
(5) PB Age First Job GT 31 And Earlier Begin (e)
(6) Inconsistent Info (h)
(7) PB N.W.,Age First Job Not Answ (f).
(8) PB N.W.,Age First Job,Spell N.A. (g)
(9) Completely Missing
(10) SP No Info In LELA And PBIOSPE (i)
(11) YOUTH-files (y)
(12) PB Empl. And Earlier Begin (y)
(13) N.W. And Later Begin (y)
(14) Inval. Info And Later Empl. (y)
(15) Empl. N.A. And Spell (y)
(16) Completely Missing
(17) Inconsistent Info (y)
(18) SP No Info YOUTH And PBIOSPE (y)
Variable format 2-digit integer

Comment AGEINFO is a pointer variable indicating the source of the age information for AGEFJOB. For ***lela***-respondents, the following coding procedure is applied:

a) For people who are or have ever been employed at the time of answering the biographical questions their age at the time of entry into the working force is taken from the ***lela***-files.

b) When we observe, that the person has not been in the working force at the time of responding, but starts to work later on, data of the ***pbiospe***-file is used. Using the spell information in ***pbiospe***, we are able to collect the age at the first job.

c) A replacement of the ***lela***-data takes place, when respondents state that they have worked before the age of fifteen, but have a spell entry later than the age of fifteen. This rule is not applied when the spell starts at the age of fifteen, since this is the minimum value for spell data in the questionnaires.

d) The same procedure is applied, when people answer, that they have never worked at the time of the interview, but have a spell which starts before the first interview.

e) In some cases the AGEFJOB value is higher than the start of the corresponding working spell in ***pbiospe***. In general, the AGEFJOB value is maintained. Only when the value is greater than 27, is it replaced by the ***pbiospe*** data. (95% of these cases have an AGEFJOB below 27.)

f) If we observe item non response concerning AGEFJOB and NOJOB, but spell information is available, the missing value is replaced by the corresponding ***pbiospe*** spell data.

g) If even the 'What did you do since you were 15' question had not been answered, there still was a chance to extract similar information out of the ***pbiospe***-file by considering the question 'What did you do every month last year'.

h) If we still had no valid information, the value of AGEFJOB was left out of the dataset.

i) Due to the fact that ***pbiospe*** information are collected only until the end of the year preceding the actual wave (in this version of ***biojob***: December 2011), for respondents without first job information from both the biography questionnaire and ***pbiospe*** we further look for a first job using information from the current wave individual questionnaire.

As described above, ***biojob*** does not contain information from the Youth Questionnaire therefore there are no observations for values (11) – (18).

NOJOB

Variable label	**"Never been employed"**
Value labels	NOJOB (1) Yes
Variable format	1-digit integer
Comment	The underlying question for the variable NOJOB is 'I have never been employed up to this date'. If NOJOB has a missing value, in general there should exist AGEFJOB information, for special cases, see above.

STILLFJ

Variable label	**"Still Employed In First Job"**
Value labels	STILLFJ (1) Yes (2) No
Variable format	1-digit integer
Comment	This variable is based on the question 'Are you still employed in the same job and at the same place?'. It applies only to ***lela***-respondents who do not state 'I have never been gainfully employed.

FULLTIME

Variable label	**"First Job Full Time"**
Value labels	FULLTIME (1) Full time job (2) Part time job
Variable format	1-digit integer
Comment	The FULLTIME-variable is used to indicate, whether the first job of a person was a full-time or a part-time job. This variable is generated out of the ***pbiospe***-file for all respondents. For persons with first job information stemming from the Biography Questionnaires, FULLTIME possibly does not refer to the declared first job if ***pbiospe*** does not contain the respective job spell (i.e. due to item non response or incomplete answering of the activity biography within the Biography Questionnaire).

OCCFJOB

Variable label	**"Occ. Position First Job"**
Value labels	OCCFJOB (1) Blue Collar W. (2) Self-Employed (3) White Collar W. (4) Civil Servant
Variable format	1-digit integer
Comment	The variable OCCFJOB provides information on the occupational position at the first job. The group 'Farmers' is included in the blue collar worker group. Due to the fact that the ***pbiospe***-file is used for the coding of AGEFJOB in certain cases (see above) there is less information on OCCFJOB than on AGEFJOB.

FJBLUE

Variable label	**"First Job Blue Collar"**
Value labels	FJBLUE (1) Unskilled Worker (2) Semiskilled Worker (3) Skilled Worker
Variable format	1-digit integer
Comment	The FJBLUE variable provides detailed information on the first occupational status if the person was a blue collar worker.

FJSELFE

Variable label	**"First Job Self Employed"**
Value labels	FJSELFE (1) Independent Farmer

(2) Freelance
(3) Other Self Employed
(4) Within Family Business

Variable format 1-digit integer

Comment The FJSELFE variable provides detailed information on the first occupational status if the person was self-employed.

FJSEFSIZ

Variable label **"No. of Employees First Job Self Employed"**
Value labels FJSEFSIZ
(1) No employees
(2) <=9 employees
(3) >=10 employees
Variable format 1-digit integer

Comment FJSEFSIZ gives the number of employees in the respondent's firm.

FJWHITE

Variable label **"First Job White Collar"**
Value labels FJWHITE
(1) Unskilled Labour, Without Degree
(2) Unskilled Labour, With Degree
(3) Skilled Labour
(4) Professional Labour
Variable format 1-digit integer

Comment FJWHITE gives detailed information on persons, who were first employed as white collar workers.

FJCIVS

Variable label **"First Job Civil Servant"**
Value labels FJCIVS
(1) Low Level Civil Servant
(2) Middle Level Civil Servant
(3) High Level Civil Servant
(4) Executive Civil Servant
Variable format 1-digit integer

Comment FJCIVS provides detailed information on first employment as a public servant

STBA

Variable label **"StaBua Vocational Classification First Job"**
Value labels STBA

Variable format 4-digit integer

Comment STBA provides information on the StaBua Vocational Classification of the first job. STBA builds on information from the Biography Questionnaire. Respondents answer the question on their first occupational title in their own words, and this response is entered into a blank in the questionnaire. Due to data protection regulations, this information cannot be provided to data users and was therefore completely recoded by Infratest Sozialforschung. For detailed information on the generation and coding of STBA see the documentation on ***pgen***.

ISCO88

Variable label **"4 Digit ISCO-88 Occupation Code First Job"**
Value labels ISCO88
Variable format 4-digit integer

Comment ISCO88 provides information on the ISCO-88 Occupation Code of the first job. ISCO88 builds on information from the Biography Questionnaire. Respondents answer the question on their first occupational title in their own words, and this response is entered into a blank in the questionnaire. Due to data protection regulations, this information cannot be provided to data users and was therefore completely recoded by Infratest Sozialforschung. For detailed information on the generation and coding of ISCO88 see the documentation on ***pgen***.

EGP

Variable label **"EGP Class Category ISCO-88 First Job"**
Value labels EGP
Variable format 2-digit integer

Comment EGP provides information on the EGP Class Category of the first job. For detailed information on the generation and coding of EGP see the documentation on ***pgen***.

ISEI

Variable label **"Ganzeboom ISEI-Status88 First Job"**
Value labels ISEI
Variable format 2-digit integer

Comment ISEI provides information on the ISEI Class Category of the first job. For detailed information on the generation and coding of ISEI see the documentation on ***pgen***.

MPS

Variable label	**"Magnitude Prestige Scale First Job"**
Value labels	MPS
Variable format	5-digit real
Comment	MPS provides information on the Magnitude Prestige Scale of the first job. For detailed information on the generation and coding of MPS see the documentation on ***pgen***.

SIOPS

Variable label	**"Treiman Standard Int. Occ. Prestige First Job"**
Value labels	SIOPS
Variable format	2-digit integer
Comment	SIOPS provides information on the Treiman Standard International Occupational Prestige Scale of the first job. For detailed information on the generation and coding of SIOPs see the documentation on ***pgen***.

CIVSFJ

Variable label	**"First Job In Civil Service"**
Value labels	CIVSFJ (1) Yes
Variable format	1-digit integer
Comment	CIVILSFJ indicates if the first job was assigned to the civil service or not. Note that this variable only applies to public servants (German: Beamte) and not to other employees in the public sector.

CURREMPL

Variable label	**"Employed At Time Of Bio Interview"**
Value labels	CURREMPL (1) Yes (2) No
Variable format	1-digit integer
Comment	This variable is based on the question 'Are you gainfully employed at the current time?'. The question applies only to ***$lela*** respondents who do not state 'I have never been gainfully employed' or 'Still employed in the first job'.

YEARLAST

Variable label	**"Year Of Last Employment"**
Variable format	4-digit integer

Comment This variable is based on the question 'When was the last time you were gainfully employed?'. The question applies only to ***$lela*** respondents who do not make at least one of the following statements in their biography interview:
'I have never been gainfully employed.'
'Still employed in the first job'
'Gainfully employed at the current time'.

SCOPELJ

Variable label **"Last Job Full-/Part-Time"**
Value labels SCOPELJ
(1) FT Employed
(2) PT Employed
(3) Marg./Irreg.Empl.
Variable format 1-digit integer

Comment SCOPELJ indicates if the last job was a full time or part time job. Information is only provided for respondents who answer the respective question within the Biography Questionnaires. The respective question applies only to respondents who do not make at least one of the following statements:
'I have never been gainfully employed.'
'Still employed in the first job'
'Gainfully employed at the current time'.

CIVILSLJ

Variable label **"Last Job In Civil Service"**
Value labels CIVILSLJ
(1) Yes
(2) No
Variable format 1-digit integer

Comment CIVILSLJ indicates if the last job was assigned to the civil service or not. Information is only provided for respondents who answer the respective question within the Biography Questionnaires. The respective question applies only to respondents who do not make at least one of the following statements:
'I have never been gainfully employed.'
'Still employed in the first job'
'Gainfully employed at the current time'.

NACELJ

Variable label **"2 Digit NACE Industry,Sector (Last Job)"**
Value labels NACELJ
Variable format 3-digit integer

Comment NACELJ provides information on the NACE Industry Code on the industry sector the respondent was employed in during the last job.

Respondents answer the question in their own words regarding the industry in which they are currently working, and this response is entered into a blank in the questionnaire. For detailed information on the generation and coding of NACELJ see the documentation on ***pgen***.

OCCLJOB

Variable label **"Occ. Position Last Job"**
Value labels OCCLJOB
(1) Blue Collar W.
(2) Self-Employed
(3) White Collar W.
(4) Civil Servant
Variable format 1-digit integer

Comment The variable OCCLJOB provides information on the occupational position at the last job. Information is only provided for respondents who answer the respective question within the Biography Questionnaires. The respective question applies only to respondents who do not make at least one of the following statements:

'I have never been gainfully employed.'
'Still employed in the first job'
'Gainfully employed at the current time'.

LJBLUE

Variable label **"Last Job Blue Collar"**
Value labels LJBLUE
(1) Unskilled Worker
(2) Semiskilled Worker
(3) Skilled Worker
Variable format 1-digit integer

Comment The LJBLUE variable provides detailed information on the last occupational status if the person was a blue collar worker.

LJSELFE

Variable label **"Last Job Self Employed"**
Value labels LJSELFE
(1) Independent Farmer
(2) Freelance
(3) Other Self Employed
(4) Within Family Business

Variable format 1-digit integer

Comment The LJSELFE variable provides detailed information on the last occupational status if the person was self-employed.

LJSEFSIZ

Variable label	**"No. of Employees Last Job Self Employed"**
Value labels	LJSEFSIZ (1) No employees (2) <=9 employees (3) >=10 employees
Variable format	1-digit integer
Comment	LJSEFSIZ gives the number of employees in the respondent's firm.

LJWHITE

Variable label	**"Last Job White Collar"**
Value labels	LJWHITE (1) Unskilled Labour, Without Degree (2) Unskilled Labour, With Degree (3) Skilled Labour (4) Professional Labour
Variable format	1-digit integer
Comment	LJWHITE gives detailed information on persons, who were last employed as white collar workers.

LJCIVS

Variable label	**"Last Job Civil Servant"**
Value labels	LJCIVS (1) Low Level Civil Servant (2) Middle Level Civil Servant (3) High Level Civil Servant (4) Executive Civil Servant
Variable format	1-digit integer
Comment	LJCIVS provides detailed information on last employment as a public servant.

Documentation *hhrf* and *phrf*

Calculating person and household level weights

Rainer Siegers

Note that this documentation is in German. Translations will be provided if necessary.

Inhalt

Hochrechnung in „Familien in Deutschland“

„Familien in Deutschland“ (FiD) ist angelegt als Erweiterung der (Mikro-)Datenbasis für Haushalte, die in bestimmten Familienkonstellationen leben. Die Konzeption der Studie erfolgte sowohl im Hinblick auf die Erhebungsinstrumente als auch bei der Form der Datenaufbereitung in Anlehnung an das „Sozio-oekonomische Panel“ (SOEP). FiD ist hier als eine Ergänzungsstichprobe zu verstehen, die zusätzliche Beobachtungen für familienpolitisch und -wissenschaftlich relevante Teilpopulationen liefert. Beide Datensätze bieten zusammengenommen ausreichend hohe Fallzahlen für differenzierte Analysen unterschiedlicher Familientypen. Im Einzelnen handelt es sich bei den Familientypen in FiD um Niedrigeinkommenshaushalte mit Kindern, Alleinerziehendenhaushalte, Haushalte mit mindestens drei minderjährigen Kindern sowie Haushalte in denen Kleinkinder der Geburtskohorten 2007 bis 2010 leben. Als Alleinerziehendenhaushalte sind die Haushalte definiert, in denen genau eine volljährige Person mit mindestens einer minderjährigen Person zusammenlebt. Niedrigeinkommenshaushalte werden anhand einer Einkommensgrenze definiert, die je nach Haushaltszusammensetzung variiert. Bei Alleinerziehenden liegt diese Grenze bei einem monatlichen Haushaltsnettoeinkommen von bis zu 1500 €, bei Familien mit einem Kind bei bis zu 2000 € und bei Familien mit mehr als einem Kind bei bis zu 2500 €. Haushalte der Kohorten sind solche, in denen sich mindestens ein Kind befindet, das in den Jahren 2007 bis (einschließlich März) 2010 geboren wurde. Die Vorgehensweise zur Hochrechnung der FiD-Daten orientiert sich grundsätzlich an den im SOEP praktizierten Hochrechnungsverfahren - unter Berücksichtigung FiD-spezifischer Umstände beim Design der Studie.

Gewichtungsansatz

Die Nutzung von Hochrechnungsfaktoren basiert auf der Idee, dass jeder beobachtete Fall der Stichprobe einen Teil der Grundgesamtheit repräsentiert. Grundgesamtheit von FiD sind alle Haushalte in Deutschland, die zum Zeitpunkt der Stichprobenziehung in einer der oben genannten Familienkonstellationen leben sowie sämtliche Personen in diesen Haushalten. Im Idealfall einer einfachen Zufallsziehung von Haushalten und der realisierten Teilnahme jedes einzelnen Haushaltes würde jeder Fall die gleiche Anzahl an Haushalten repräsentieren und alle den gleichen Hochrechnungsfaktor bekommen, der dem Kehrwert seiner Ziehungswahrscheinlichkeit entspricht. Tatsächlich führen sowohl das Ziehungsdesign als auch Unit-Nonresponse, also die Verweigerung eines Interviews durch Befragte, zu

Unterschieden in der Wahrscheinlichkeit, Teil der Stichprobe zu sein. Ziel der Hochrechnung ist die Schätzung des Kehrwerts dieser Wahrscheinlichkeit für jede Erhebungseinheit.
Das Hochrechnungsverfahren besteht dabei grundsätzlich aus drei Schritten:

1. Ermittlung der Design-Gewichte: Die Designgewichte bilden die unterschiedlichen Ziehungswahrscheinlichkeiten ab, die bereits durch das Ziehungsdesign festgelegt sind. Konkret handelt es sich um die inverse Wahrscheinlichkeit, Teil der Brutto-Stichprobe zu sein. Diese Bruttostichprobe wird dann vom Erhebungsinstitut zur Befragung kontaktiert. Die Ziehungswahrscheinlichkeiten variieren, wenn beim Design der Stichprobe bereits systematische Unterschiede nach bestimmten Kriterien bestehen. Bei FiD werden beispielsweise die Haushaltstypen mit unterschiedlichen Wahrscheinlichkeiten gezogen. Darüber hinaus findet in der Kohorten-Stichprobe eine überproportionale Ziehung von Haushalten mit Migrationshintergrund statt.
2. Schätzung der Unit-Nonresponse-Gewichte: Nicht alle der für eine Befragung vorgesehenen Haushalte nehmen tatsächlich teil. Die Ausfälle auf dem Weg der ausgewählten Haushalte zur tatsächlichen Befragung können als weitere zufällige und von den vorangegangenen Schritten unabhängige Ziehungsstufen interpretiert werden, deren inverse Teilnahmewahrscheinlichkeit geschätzt werden kann. Damit kann auf Basis beobachteter Eigenschaften dafür kontrolliert werden, dass Haushalte z.B. unterschiedliche Verweigerungsquoten aufweisen. Die Gewichte der unabhängigen Ziehungsstufen werden miteinander multipliziert und ergeben das Gewicht, das der inversen Wahrscheinlichkeit entspricht, alle Stufen durchlaufen zu haben.
3. Kalibrierung der Gewichte: Schließlich werden die geschätzten Hochrechnungsfaktoren unter Zuhilfenahme externer Informationen über ausgewählte Randverteilungen der Grundgesamtheit kalibriert. Im Fall von FiD wurden die „tatsächlichen“ Randverteilungen unter Vernachlässigung eventueller Saisoneffekte aus dem Mikrozensus übernommen.

Stichproben in FiD

Die Haushalte des FiD-Datensatzes wurden in drei voneinander unabhängigen Stichproben gezogen. Die praktische Umsetzung des Ziehungsdesigns und die Feldarbeit wurden von TNS Infratest durchgeführt.

Bei der ersten 2010 gezogenen Stichprobe handelt es sich um eine sogenannte „Screening-Stichprobe“ zum Sampling von Niedrigeinkommenshaushalten mit Kindern, Alleinerziehendenhaushalten und Haushalten mit mindestens drei minderjährigen Kindern, kurz „Sreening-Stichprobe 2010“ genannt. Der Name Screening-Stichprobe beruht darauf, dass vor dem bei FiD üblichen persönlichen Interviewerkontakt ein telefonisches Screening in Frage kommender Haushalte stattgefunden hat, um vorab die Zugehörigkeit zu einer der genannten Zielpopulationen sowie die grundsätzliche Teilnahmebereitschaft des Haushaltes zu klären. Dieses Verfahren wurde gewählt, da weniger als 10% der deutschen Haushalte in eine der Zielkategorien fallen und der erhebliche Mehraufwand von Face-to-Face-Kontakten in diesem Fall nicht gerechtfertigt gewesen wäre. Als Basispopulation dienten Befragte sogenannter Omnibus-Umfragen (fortan Bus-Umfragen oder Busse genannt) von TNS Infratest aus den 18 Monaten vor dem Screening. Bei den Bus-Umfragen handelt es sich um regelmäßig von TNS Infratest durchgeführte Mehrthemenbefragungen, die teilweise telefonisch und teilweise im Face-to-Face-Interview durchgeführt werden und bevölkerungsrepräsentativ angelegt sind.

Bei der zweiten 2010 gezogenen Stichprobe wurden Haushalte mit Kindern aus den Geburtskohorten der Jahre 2007-2010 („Kohorten-Stichprobe“) mit Hilfe von Informationen der Einwohnermeldeämter in einem zweistufigen Verfahren gezogen. Auf der ersten Stufe wurden Einwohnermeldeämter selektiert, in denen dann gemeldete Kinder der vier Kohorten zufällig ausgewählt wurden. Deren Haushalte wurden durch TNS Infratest kontaktiert und zur Teilnahme an der Studie animiert.

Die dritte Stichprobe wurde 2011 als Ergänzung zu den Alleinerziehendenhaushalten und Haushalten mit mindestens drei minderjährigen Kindern, jedoch nicht der Niedrigeinkommenshaushalte, nach demselben Verfahren wie 2010 gezogen und zur FiD-Population hinzugefügt. Entsprechend wird diese Stichprobe als „Screening-Stichprobe 2011“ bezeichnet. Eine detailliertere Beschreibung des Ziehungsverfahrens der Stichproben findet sich in den Methodenberichten von TNS Infratest[16].

[16] Jänsch A., Huber S., Siegel N. A., Stimmel S., Geue D.: *„Familien in Deutschland“ - (FiD) 2010 - Methodenbericht über Anlage und Ergebnisse der FiD-Stichproben* München 2011.

Die konkreten Gewichtungsschritte zur Hochrechnung des FiD-Datensatzes vom Design-Gewicht bis zur Kalibrierung sowie zur Integration der Stichproben werden im Folgenden beschrieben. Das langfristige Hochrechnungskonzept von FiD wird sowohl die separate Analyse des FiD-Datensatzes als auch die eines von SOEP und FiD integrierten Datensatzes ermöglichen.[17] Die integrierten Hochrechnungsfaktoren bedingen die Gewichtungsfaktoren des SOEP und damit abgeschlossene SOEP-Daten. Da die FiD- und SOEP-Daten einer Welle zu unterschiedlichen Zeitpunkten bereitgestellt werden, enthält die erste Version einer FiD-Welle keine Variablen zur gemeinsamen Hochrechnung der letzten Welle. Diese werden bereitgestellt, nachdem auch die SOEP-Daten der betreffenden Welle veröffentlicht sind.

Hochrechnung der Screening-Stichproben 2010 und 2011

Abbildung 1 veranschaulicht den Auswahlprozess der Screening-Stichprobe 2010 bis hin zu den tatsächlich im Datensatz enthaltenen Netto-Haushalten. Die Realisierung der Screening-Stichprobe 2011 erfolgt nach dem identischen Ablauf, jedoch mit geringeren Fallzahlen. Im Gewichtungsprozess spiegeln sich die einzelnen Schritte im Wesentlichen wider. Die Schätzungen der Wahrscheinlichkeiten, die einzelnen Gewichtungsstufen zu „überstehen", basieren jeweils auf Modellen einer logistischen Regression. Insgesamt werden bei der Hochrechnung der Screening-Stichproben sieben Gewichtungsschritte durchlaufen, deren Modelle im Folgenden kurz erläutert werden (eine Übersicht der Modelle findet sich im Anschluss an deren Erläuterung in Tabelle 1).

Ausgangssituation: Die Screening-Stichprobe 2010 beruht auf Personen, die bei TNS Infratest an einem „Face-to-Face-Bus" (auch F2F-Bus) der letzten 18 Monate oder einem „Telefon-Bus" der letzten 12 Monate vor dem Zeitpunkt des Screenings im Januar 2010 teilgenommen haben. Die Screening-Stichprobe 2011 beruht auf einem „Face-to-Face-Bus" bzw. einem „Telefon-Bus" der letzten 12 Monate vor dem Zeitpunkt des Screenings im Januar 2011. Die zur Stichprobenauswahl benötigten Informationen über Haushaltszusammensetzung und -einkommen sind demnach bis zu 18 Monate alt. Beide Busse sind repräsentativ für Personen ab 14 Jahren in Privathaushalten in Deutschland, mit der Einschränkung, dass der F2F-Bus sich auf deutsche Staatsangehörige beschränkt, der Telefon-Bus auf deutschsprachige Personen mit Festnetzanschluss. Für die erneute Kontaktaufnahme der Bus-Teilnehmer zur Überprüfung, ob sie Teil der Zielpopulation von FiD sind, ist es notwendig, dass sie ihre

---

Geue D. ., Siegel N., Konhäuser L.: *„Familien in Deutschland" - (FiD) 2011- Methodenbericht über Anlage und Ergebnisse der FiD- Stichproben* München 2011.

Bereitschaft für weitere Befragungen erklärt haben. Es besteht also ein Selektionseffekt darin, nur wiederbefragungswillige Personen in der Ausgangsstichprobe zu haben. Über das Ausmaß dieser Einschränkungen stehen keine genauen Informationen zur Verfügung. Daher erlangt bei den Screening-Stichproben der abschließende Kalibrierungsschritt eine stärkere Bedeutung als üblich. Migranten-Haushalte können aufgrund der Konzentration der Busse auf Deutsche bzw. deutschsprachige Personen in der Screening-Stichprobe nicht repräsentativ erfasst werden und werden bei der Gewichtung nicht gesondert berücksichtigt.

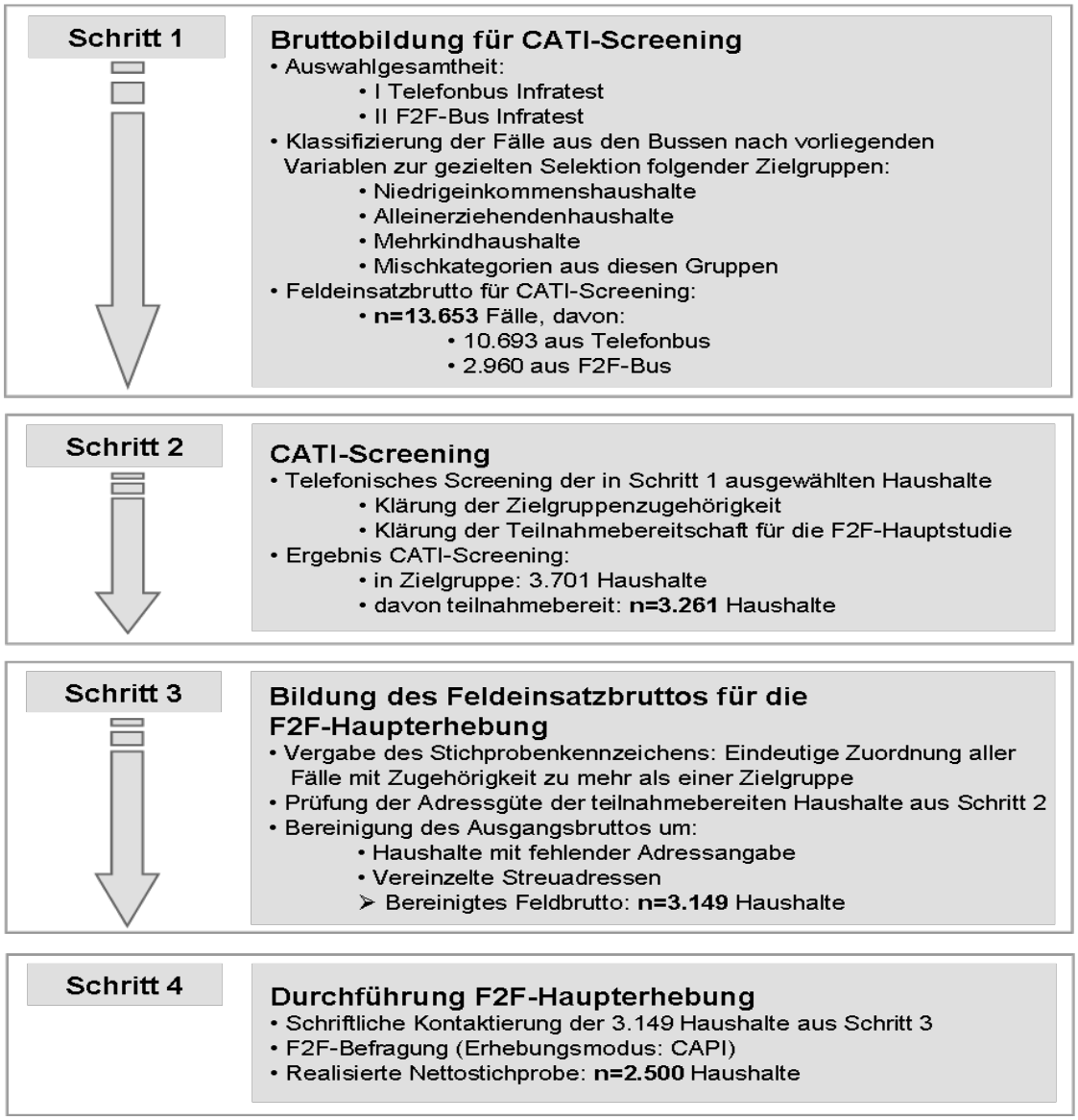


Abbildung 1: Stichprobendesign für die Screening-Stichprobe 2010, Quelle: Methodenbericht TNS Infratest

[17] Unter der „Integration“ ist zu verstehen, dass die unterschiedlichen Stichproben (bzw. Datensätze) mit Hochrechnungsfaktoren versehen werden, die eine gemeinsame Nutzung ermöglichen.

Modell 1: Aus den drei Haushaltstypen der FiD-Zielpopulation und deren Kombinationen untereinander lassen sich in der Screening-Stichprobe 2010 sieben disjunkte Kategorien von Haushaltstypen bilden. Wegen der Vorgaben zu Netto-Fallzahlen der drei Haushaltstypen (1000 Niedrigeinkommenshaushalte, 500 Alleinerziehendenhaushalte, 500 Mehrkindhaushalte) variieren die Ziehungswahrscheinlichkeiten zwischen den Kategorien. Abbildungen 2a und 2b zeigen für die Screening-Stichproben 2010 und 2011 die unterschiedlichen Verteilungen der finalen Hochrechnungsfaktoren auf Haushaltsebene nach Haushaltstypen für das erste Befragungsjahr. Die dort verdeutlichten Unterschiede unterstreichen die Relevanz der Hochrechnungsfaktoren bei Analysen mit FiD. Die Zugehörigkeit zu den Kategorien von Haushaltstypen wurde zunächst von TNS Infratest auf Basis der Bus-Informationen festgelegt. Beim ersten Modell der Gewichtung wird die Wahrscheinlichkeit geschätzt, aus den in Frage kommenden Bus-Personen für das telefonische Screening ausgewählt zu werden. Als erklärende Variablen werden neben den sieben Haushaltstypen der Bildungsstand, das Geschlecht und der Erwerbsstatus der befragten Person genutzt. Für die Screening-Stichprobe 2011 ergeben sich aufgrund der Merkmale nur drei disjunkte Haushaltstypen, die in ein ansonsten identisches Modell aufgenommen werden.

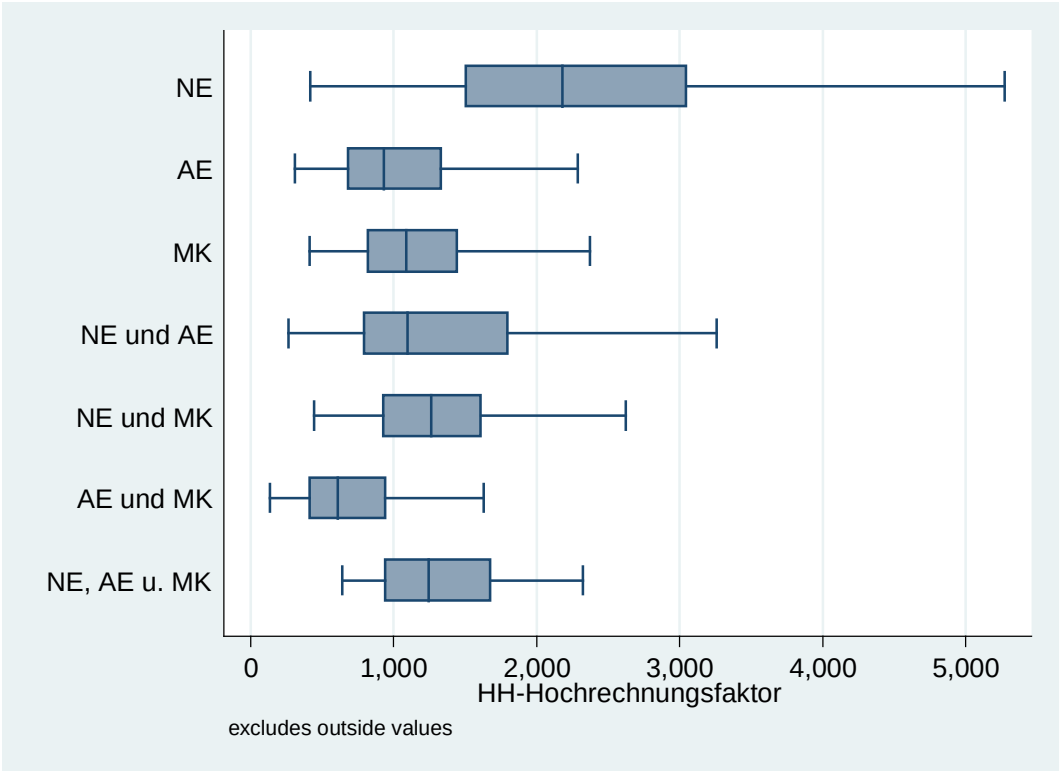


Abbildung 2a: Boxplots der Hochrechnungsfaktoren der Screening Stichprobe 2010 nach Haushaltstypen (NE: Niedrigeinkommenshaushalte, AE: Alleinerziehendenhaushalte, MK: Mehrkindhaushalte)

Modell 2: Nicht alle der für das telefonische Screening ausgewählten Personen können erneut kontaktiert werden. Es wurden 13.653 Personen kontaktiert, von denen 10.396 telefonisch erreicht werden konnten (2011: 8.400 kontaktiert, 6.461 erreicht). Das zweite Modell schätzt die Wahrscheinlichkeit, telefonischen Kontakt zu den ausgewählten Personen herstellen zu können. Hierzu werden neben den Einkommensangaben sowie den demografischen und räumlichen Indikatoren, die aus dem letzten Bus-Kontakt bekannt sind, zusätzliche Informationen von INKAR[18] verwendet. Hierbei handelt es sich um soziodemographische und räumliche Informationen auf Ebene der Gemeinde, in der die Person zuletzt kontaktiert wurde. Diese Wahrscheinlichkeit wird für die disjunkten Haushaltstypen (2010: sieben; 2011: drei) in einem gemeinsamen Modell geschätzt, jedoch werden Interaktionseffekte aller erklärenden Variablen mit den Haushaltstypen zugelassen, so dass spezifische Effekte auf die Kontaktwahrscheinlichkeit einzelner Haushaltstypen berücksichtigt werden.

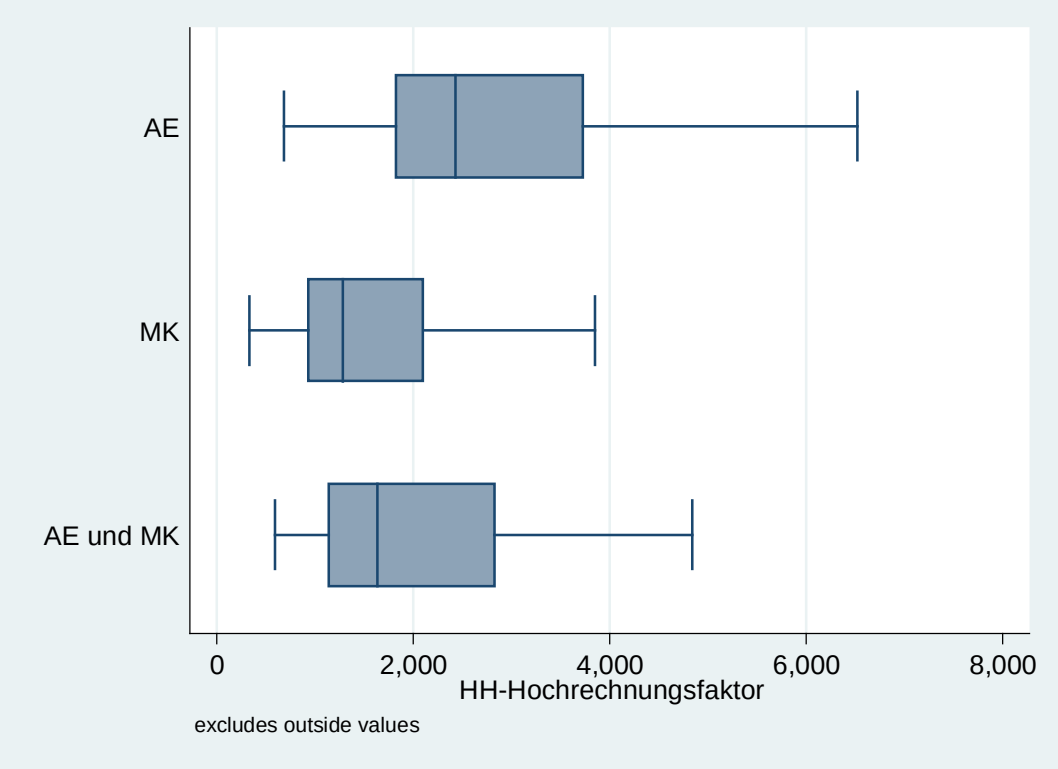


Abbildung 2b: Boxplots der Hochrechnungsfaktoren der Screening Stichprobe 2011 nach Haushaltstypen (AE: Alleinerziehendenhaushalte, MK: Mehrkindhaushalte)

Modell 3: Unter den kontaktierten Haushalten des Telefon-Screenings gibt es solche, die die Teilnahme am Screening verweigerten. Von den 10.396 kontaktierten Personen in 2010

[18] für weitere Informationen über die zur Verfügung stehenden Variablen siehe http://www.bbr.bund.de/nn_23470/BBSR/DE/Veroeffentlichungen/INKAR/inkar__node.html?__nnn=true

erklärten sich 6.656 Personen zum Screening-Interview bereit (2011: 4.827 von 6.461). Zur Berücksichtigung dieser Ausfälle wurde ein Modell zur Schätzung der Wahrscheinlichkeit der Teilnahmebereitschaft am Telefon-Screening erstellt. Auch hier wurden die individuellen Bus-Informationen und die wohnortspezifischen INKAR-Informationen genutzt und ein gemeinsames Modell für die gesamte Screening-Stichprobe ggf. unter Berücksichtigung von Interaktionseffekten zwischen Haushaltstypen und erklärenden Variablen erstellt.

Über die Teilnehmer am Screening-Verfahren wurden nun aktualisierte Informationen über Haushaltszusammensetzung, Einkommen, Bildung, Erwerbsstatus, Wohnort etc. erhoben. Auf Basis dieser Informationen konnte eine neue Einteilung der Haushalte in die sieben (bzw. drei) Haushaltstypen der Zielpopulation vorgenommen werden. Dabei stellte sich in 2010 heraus, dass 2.955 von 6.656 Screening-Personen nicht in Haushalten leben, die zur Zielpopulation der FiD-Screening-Stichprobe gehören (2011: 3.515 von 6.461). Erst bei Personen, die das Screening durchlaufen haben, konnte die Information für die Zuordnung zur Zielpopulation nach den von FiD vorgegebenen Kriterien bezüglich Haushaltkomposition und Einkommensgrenzen stattfinden. Die vorangegangene Auswahl aus den Infratest-Bussen musste daher auf einer großzügigeren Auswahl der At-Risk-Population beruhen. Haushalte, die sich auf dieser Stufe als nicht zur Zielpopulation gehörig herausstellten, wurden nicht weiter verfolgt und fallen daher aus dem Gewichtungsverfahren heraus. Die Entscheidung, diese Haushalte nicht weiter bei der Gewichtung zu berücksichtigen, wurde unter der Annahme getroffen, dass sie sich auf den vorangegangenen Gewichtungsschritten hinsichtlich der Ausfallwahrscheinlichkeiten nicht wesentlich von der Zielpopulation unterscheiden.

Modell 4: In 2010 erklärten sich von den 3.701 Haushalten, die das Screening durchlaufen haben und zur Zielpopulation gehören, 3.261 Haushalte grundsätzlich bereit, bei der FiD-Studie teilzunehmen. 51 Haushalte wurden als Streuadressen nicht kontaktiert. Sie wurden bei der Gewichtung als qualitätsneutrale Ausfälle gewertet und nicht weiter berücksichtigt. 61 Haushalte gaben falsche bzw. fehlerhafte Adressen an und wurden bei der Gewichtung als heimliche Verweigerer aufgefasst. Für 2010 fließen in Modell 4 also 3650 Fälle ein.
In 2011 sind es 1.154 von den 1.312 zur Zielpopulation gehörigen Haushalten, die auch an der Haupterhebung teilnehmen wollen. Hier werden 40 Streuadressen nicht kontaktiert, sowie ein Haushalt mit falschen bzw. fehlerhaften Adressen als Verweigerer eingeordnet, so dass Modell 4 in 2011 mit 1.272 Fällen berechnet wird.

Für die Haupterhebung 2010 gibt es ein Ausgangs-Brutto von 3.149 Haushalten (2011: 1.154), die sich zur Teilnahme bereit erklärt haben. Auf dieser Modellstufe wurde die Wahrscheinlichkeit zur Teilnahmebereitschaft an der Haupterhebung geschätzt, diesmal allerdings auf Basis der aktualisierten individuellen Informationen, erneut ergänzt durch Gemeindeinformationen von INKAR.

Modell 5: Die im Ausgangs-Brutto befindlichen 3.149 Haushalte in 2010 bzw. 1.154 Haushalte in 2011 werden von Infratest kontaktiert und zur Teilnahme an der Haupterhebung eingeladen. Es gibt Haushalte, die nicht aufgefunden werden konnten oder zu denen aus anderen Gründen kein Kontakt zustande kam. Dies betrifft allerdings nur wenige Haushalte, so dass auf ein eigenes Modell verzichtet wird. Häufiger ist eine Ablehnung der Teilnahme an der Studie bei den kontaktierten Haushalten. Die Wahrscheinlichkeit zur Teilnahme der Brutto-Haushalte an der Befragung wird daher in einem Modell geschätzt, das beide Ausfalltypen, d.h. fehlender Kontakt und Verweigerung, gemeinsam erfasst. Da TNS Infratest die Adressen der Brutto-Haushalte bekannt sind, kann bei der Gewichtung eine weitere, kleinräumlichere externe Datenquelle genutzt werden, um genauere Informationen bezüglich der verweigernden Haushalte zu erlangen. Hierbei handelt es sich um Daten von MICROM[19], die Konsumenteninformationen auf Haus-, Straßenzug- und Marktzellenebene bereitstellen und die den Brutto-Haushalten über die Adressen zugespielt werden können. Mithilfe dieser Daten, der INKAR-Daten, den Angaben aus dem Screening-Interview und den von den Interviewern erhobenen Wohnumfelddaten wird die Teilnahmewahrscheinlichkeit der Haushalte geschätzt.

Modell 6: Das Produkt aus den Design-Hochrechnungsfaktoren und dem Kehrwert der Wahrscheinlichkeit aus Modell 5 bildet das Eingangsgewicht für den abschließenden Kalibrierungsschritt der Haushaltshochrechnungsfaktoren. Vorab wird dieses Eingangsgewicht noch modifiziert, um extreme Ausreißer zu vermeiden. Hierzu wird das Gewicht zunächst auf einen Mittelwert von 10 normiert und ab einer kritischen Grenze des

[19] Die Verknüpfung der MICROM-Daten mit den FiD-Haushalten verläuft analog zum SOEP. Zur Verknüpfung des SOEP mit den MICROM-Daten siehe Goebel J., Spieß C. K., Witte N. R. J. Gerstenberg S.: [http:/ / www.diw.de/ documents/ publikationen/ 73/ diw_01.c.78103.de/ diw_datadoc_2007-026.pdf] *Die Verknüpfung des SOEP mit Microm-Indikatoren: Der MICROM-SOEP Datensatz. Berlin*. 2007.
Weitere Informationen zu MICROM unter http://www.microm-online.de/Deutsch/Microm/index.jsp

2,5-fachen des Medians logarithmisch korrigiert.[20] Angepasst werden die Gewichte schließlich in einem iterativen Verfahren an Ränder des Mikrozensus, die vom Statistischen Bundesamt zur Verfügung gestellt wurden.[21] Im Einzelnen handelt es sich dabei um die Randverteilungen der Bundesländer, Gemeindegrößenklassen, Bezug von Arbeitslosengeld II, Höhe des Haushaltseinkommens, Haushaltsgröße, Haushaltskomposition (Anzahl der Kinder verschiedener Altersklassen), der die FiD-Zielpopulation definierenden Haushaltstypen sowie zur späteren Integration um Überschneidungen mit den Geburtskohorten der Kohorten-Stichprobe. Aus den Angaben in der Hauptstichprobe ergibt sich, dass von den 2.500 teilnehmenden Haushalten der 2010 gezogenen Screening-Stichprobe 237 wegfallen (2011: 9 von 924 Haushalten), weil sie nicht Teil der Zielpopulation sind. Erklären lässt sich diese Reduzierung im Wesentlichen durch die genauere Klassifizierung nach der Abfrage des Haushaltseinkommens bzw. durch Änderungen der Haushaltskonstellation zwischen den Zeitpunkten des Screenings und der Haupterhebung. Außerdem erteilt in der Haupterhebung diejenige Person Auskunft über den Haushalt, die sich am besten mit ihm auskennt, während in den Bussen und dem Screening-Interview eine zufällig ausgewählte Person über 14 Jahren befragt wurde. Die Haushalte der Screening-Stichprobe 2010 außerhalb der Zielpopulation werden in der ersten Welle von FiD weitergegeben, erhalten aber einen Hochrechnungsfaktor von Null und werden in den folgenden Wellen nicht wieder befragt. Haushalte der Screening-Stichprobe 2011 außerhalb der Zielpopulation werden dagegen auch in den weiteren Wellen befragt, erhalten aber ebenfalls einen Hochrechnungsfaktor von Null.

Modell 7: Ausgehend von den kalibrierten Haushaltshochrechnungsfaktoren werden im Anschluss Hochrechnungsfaktoren auf Personenebene erstellt, die sich durch eine weitere iterative Randanpassung ergeben. Genutzt wurden die Randverteilungen für Alter, Geschlecht, Erwerbstatus, Bildung, Zugehörigkeit zu den Haushaltstypen, Überschneidungen mit den Haushalten der Kohorten-Stichprobe und einer Korrektur für Partial *Unit-Non-Response (PUNR) nach Art der Haushaltszusammensetzung. PUNR-Fälle selbst, also Personen aus teilnehmenden Haushalten, die die Teilnahme an der Befragung verweigern, bekommen ein Gewicht von Null.* Tabelle 1 zeigt eine zusammenfassende Übersicht über die einzelnen Schritte bei der Gewichtung der Screening-Stichprobe.

---

[20] nach folgender Formel $W_{korr} = \begin{cases} 2{,}5 \cdot \tilde{W}_{orig} + \ln(W_{orig} + 1 - (2{,}5 \cdot \tilde{W}_{orig})) & \text{, wenn } W_{orig} > 2{,}5 \cdot \tilde{W}_{orig} \\ W_{orig} & \text{, sonst} \end{cases}$

[21] Besonderer Dank gilt Robert Herter-Eschweiler, der uns diese recht aufwändige Sonderauswertung des Statistischen Bundesamtes zusammengestellt hat.

	Gegenstand der Schätzung	Genutzte Information	Fallzahl 2010		Fallzahl 2011	
			vor Modell-schritt	nach Modell-schritt	vor Modell-schritt	nach Modell-schritt
Modell 1	Auswahlwkt. für das Screening aus Infratest-Bussen	Eingeschränkte Informationen aus den Bus-Befragungen	23193	13653	19041	8400
Modell 2	Kontaktwahrscheinlichkeit bei Screening	Angaben aus den Bus-Befragungen sowie INKAR-Daten	13653	10396	8400	6461
Modell 3	Teilnahmewahrscheinlichkeit am Screening	Angaben aus den Bus-Befragungen sowie INKAR-Daten	10396	6656	6461	3515
Modell 4	Wkt. für Bereitschaft zur Teilnahme an Haupterhebung	Angaben aus den Bus-Befragungen, dem Screening sowie INKAR-Daten	3650	3149	1272	1154
Modell 5	Wkt. für tatsächliche Teilnahme an Haupterhebung	Angaben aus den Bus-Befragungen, dem Screening, durch Interviewer sowie INKAR- und MICROM-Daten	3149	2500	1154	924
Modell 6	Kalibrierung auf Haushaltsebene	Angaben aus dem Mikrozensus	2263	2263	915	915
Modell 7	Kalibrierung auf Personenebene	Angaben aus dem Mikrozensus	7994	7994	3548	3548

Tabelle 1: Übersicht über die einzelnen Schritte der Gewichtung der Screening-Stichproben

Hochrechnung der Kohorten-Stichprobe

Die folgende Grafik veranschaulicht den Auswahlprozess der Kohorten-Stichprobe bis hin zu den tatsächlich im Datensatz enthaltenen Netto-Haushalten. Im Gewichtungsprozess spiegeln sich die einzelnen Schritte im Wesentlichen wider. Im Folgenden werden die einzelnen Stufen im Hinblick auf die Gewichtung der Kohorten-Stichprobe betrachtet (Tabelle 2 gibt den Überblick über die einzelnen Modelle).

Modell 1: Die Kohorten-Stichprobe wurde in einem geschichteten, mehrstufigen Verfahren gezogen. In einem ersten Schritt wurden zufällig Einwohnermeldeämter gezogen. Innerhalb der Daten der Einwohnermeldeämter wurden in einem weiteren Schritt zufällig Kinder der interessierenden Geburtskohorten ausgewählt. Um dem bei Surveys bekannten Phänomen zu begegnen, dass Personen mit Migrationshintergrund deutlich geringere Responseraten als die

autochthone Bevölkerung aufweisen, wurden Migranten bei FiD systematisch überrepräsentiert erhoben. Dafür wurden zunächst beim zweiten Schritt je Kohorte mehr Adressen als benötigt gezogen. Identifiziert wurde der Migrationshintergrund über die bei den Einwohnermeldeämtern hinterlegte Information zur Nationalität. Darüber hinaus wurde Migrationshintergrund der Kinder aus den gezogenen Adressen anhand eines onomastischen Verfahrens identifiziert. Aus dem gezogenen Adresspool wurde nun in einem dritten Schritt eine vorgegebene Zahl von Kindern gezogen. Die Anzahl der identifizierten Migranten wurde anschließend für jeden Sampling-Point aus dem restlichen Adresspool verdoppelt. Trotz der bevölkerungsproportionalen Ziehung kommt es daher designbedingt zu unterschiedlichen Ausgangswahrscheinlichkeiten in der Stichprobe. Diese sind in erster Linie vom kohortenspezifischen Migrantenanteil der einzelnen Sample-Points abhängig. Die entsprechenden Designgewichte wurden von TNS Infratest zur Verfügung gestellt.

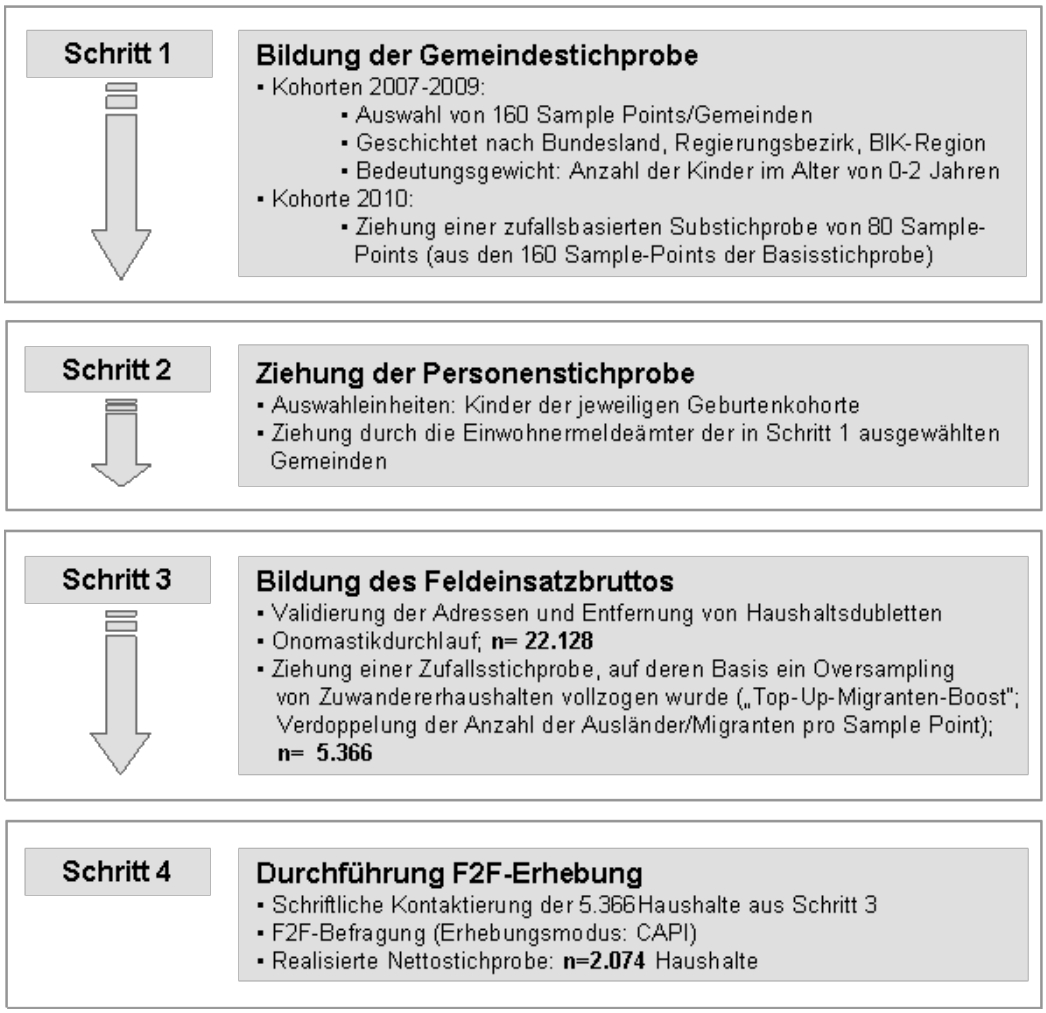


Abbildung 3: Stichprobendesign für die Kohorten-Stichprobe, Quelle: Methodenbericht TNS Infratest

Modell 2: Von den 5.366 zur Befragung vorgesehenen Haushalten nahmen schließlich 2.074 Haushalte an der Befragung teil. Die anderen Haushalte der Brutto-Stichprobe teilen sich in

604 qualitätsneutrale Ausfälle, 1.018 Ausfälle, die nicht kontaktiert werden konnten sowie 1.670 Ausfälle durch Verweigerung der Teilnahme. Im zweiten Modell wird die Kontaktwahrscheinlichkeit geschätzt. Dabei kann, wie schon bei der Hochrechnung der Screening-Stichproben, neben den von den Interviewern erhobenen Informationen auf externe Daten von INKAR und MICROM zurückgegriffen werden.

Modell 3: Im Anschluss wird die Kooperationswahrscheinlichkeit geschätzt. Hier kann auf den gleichen Variablenstamm wie in Modell 2 zurückgegriffen werden. Beide Gewichte multipliziert mit den Design-Hochrechnungsfaktoren ergeben die Ausgangshochrechnungs-faktoren für die abschließende Kalibrierung.

Modell 4: Auch die Hochrechnungsfaktoren der Kohorten-Stichprobe wurden abschließend über eine Randanpassung kalibriert. Dafür standen erneut Informationen des Mikrozensus des Statistischen Bundesamtes zur Verfügung. Die Stichprobe zur Geburtskohorte 2010 wurde bei den Einwohnermeldeämtern mit Abschluss des ersten Quartals gezogen und enthält daher (fast) nur 2010-Geborene aus diesem Zeitraum. Dies wird bei der Randanpassung berücksichtigt, so dass die hochgerechnete Stichprobe für das Jahr 2010 nur den Teil des Geburtenjahrgangs 2010 repräsentiert, der zum Zeitpunkt der Ziehung durch FiD erfasst wurde[22]. Die zur Verfügung stehenden Variablen zur Randanpassung entsprechen denen der Screening-Stichproben, ergänzt um die Häufigkeit von Mehrlingen und Kindern aus den jeweils anderen Kohorten in den Kohortenhaushalten.

Modell 5: Die Erstellung der Hochrechnungsfaktoren auf Personenebene geschieht analog zu dem Verfahren der Screening-Stichprobe (dort Modell 7).

Tabelle 2 zeigt eine Übersicht über die einzelnen Schritte bei der Gewichtung der Kohorten-Stichprobe.

	Gegenstand der Schätzung	**Genutzte Information**	**Fallzahl**	
			vor Modellschritt	**nach Modellschritt**

[22] In älteren FiD-Versionen (bis v3.0) wurden die in FiD enthaltenen Fälle der Geburtskohorte 2010 als repräsentativ für das komplette Geburtsjahr 2010 interpretiert und entsprechend hochgerechnet. Zum Unterschied zwischen FiD-Hochrechnungsfaktoren ab v3.1 und denen bis v3.0 siehe auch das entsprechende Dokument zur Weitergabe von v3.1 (November 2013) im Anhang dieser Dokumentation.

Modell 1	Oversampling von Migranten	Angaben der Einwohner-meldeämter	5366	5366
Modell 2	Kontaktwahrscheinlichkeit bei Haupterhebung	Angaben durch Interviewer sowie INKAR- und MICROM-Daten	4762	3744
Modell 3	Teilnahmewahrscheinlichkeit bei Haupterhebung	Angaben durch Interviewer sowie INKAR- und MICROM-Daten	3744	2074
Modell 4	Kalibrierung auf Haushaltsebene	Angaben aus dem Mikrozensus	2074	2074
Modell 5	Kalibrierung auf Personenebene	Angaben aus dem Mikrozensus	7670	7670

Tabelle 2: Übersicht über die einzelnen Schritte der Gewichtung der Kohorten-Stichprobe

Integration der FiD-Stichproben

Mit den bisher beschriebenen Hochrechnungsfaktoren lassen sich Analysen für die jeweilige Population durchführen, also beispielsweise für die Haushalte, die 2010 als Niedrigeinkommenshaushalte definiert wurden und sich in der Screening-Stichprobe befinden. Allerdings nutzt man dann nicht das volle Potential, weil Fälle nicht berücksichtigt werden, die aus den anderen Stichproben ebenfalls zur interessierenden Population gehören. Ein Niedrigeinkommenshaushalt mit Kleinkind könnte beispielsweise sowohl als Teil der Screening- als auch der Kohorten-Stichprobe gezogen worden sein. Die Integration der Screening- und Kohorten-Stichproben mit gemeinsamen Hochrechnungsfaktoren löst dieses Problem, so dass bei Analysen zusätzliche Fälle durch die Nutzung der anderen Stichproben gewonnen werden können.

Aufgrund der unterschiedlichen Ziehungsjahre erfolgt die Integration der einzelnen Stichproben nicht einheitlich. Für das Ziehungsjahr 2010 und die Integration der Screening-Stichprobe 2010 mit der Kohorten-Stichprobe werden als Ausgangswerte für die Integration die Gewichte dieser beiden Stichproben vor dem Kalibrierungsschritt herangezogen und gemeinsam auf der Haushaltsebene kalibriert. Da bei der Kalibrierung die verwendeten Eingangsgewichte Einfluss auf die Verteilung der resultierenden Hochrechnungsfaktoren haben, werden sie normiert, so dass die überschneidenden Populationen beider Stichproben mit dem gleichen durchschnittlichen Eingangsgewicht in den ersten Kalibrierungsschritt laufen. Die verwendeten Variablen für die Randanpassung entsprechen inhaltlich denen, die

für die stichprobenspezifischen Kalibrierungen herangezogen werden. In einem zweiten Schritt werden auf Basis der integrierten Hochrechnungsfaktoren auf Haushaltsebene die Personen-Hochrechnungsfaktoren erstellt.

Für die Screening-Stichprobe 2011 lassen sich zunächst, wie beschrieben, die Gewichte für die Population berechnen. Allerdings kann eine Integration mit den beiden in 2010 gezogenen Stichproben nicht einfach erfolgen, weil aufgrund der sich verändernden Population aus 2010 keine gemeinsamen Ränder zwischen den drei Stichproben vorliegen.[23] Entsprechend wird eine gemeinsame Nutzung aller FiD-Stichproben erst durch die Integration mit dem SOEP und die damit verbundene Hochrechnung auf die Gesamtbevölkerung möglich.

Integration von SOEP und FiD

Die inhaltliche Überlappung von FiD und SOEP ist erheblich, mehr als 80% der Variablen sind in beiden Datensätzen identisch erhoben worden. Dies legt eine gemeinsame Nutzung der Daten nahe, insbesondere dann, wenn Aussagen über die gesamte deutsche Bevölkerung getroffen werden sollen. Zu diesem Zweck werden in FiD auch Hochrechnungsfaktoren geliefert, die eine gemeinsame Hochrechnung von SOEP und FiD erlauben. Gemeinsame Hochrechnungsfaktoren können bereitgestellt werden, sobald die SOEP-Daten für die entsprechende Welle zur Verfügung stehen. Die Berechnung dieser Gewichte erfolgt über ein zweistufiges Verfahren, in dem zuerst die Haushalte im SOEP identifiziert werden, die die gleichen Merkmale tragen wie die Haushalte aus den FiD-Stichproben, also im Jahr 2010 Kinder in den vier Kohorten 2007-2010 haben oder Niedrigeinkommenshaushalte, Alleinerziehendenhaushalte oder Haushalte mit drei oder mehr minderjährigen Kinder bzw. im Jahr 2011 Alleinerziehendenhaushalte oder Haushalte mit drei oder mehr minderjährigen Kinder sind. Im zweiten Schritt werden dann alle Fälle beider Datensätze an die Randverteilungen der Gesamtbevölkerung angepasst, wobei die Startgewichte der überlappenden Populationen entsprechend dem Verhältnis der Fallzahlen in den einzelnen Datensätzen angepasst werden. Dieses Verfahren zur Integration der FiD-Stichproben in den Hochrechnungsrahmen des SOEP findet, wie bei anderen neuen SOEP-Stichproben, in der ersten Erhebungswelle statt.

Querschnittsgewichte für Daten nach 2010

[23] Die Screening-Stichprobe 2010 bildet die Populationen der Niedrigeinkommenshaushalte, der Alleinerziehenden und der Mehrkindfamilien in 2010 ab. In 2011 entspricht diese Gruppe allerdings nicht mehr notwendigerweise denselben Populationen, sondern gerade bei diesen Merkmalen wird es zum Teil erhebliche Fluktuationen geben. Entsprechend ist eine Kalibrierung mit im Mikrozensus als Querschnittsdatensatz gemessenen Merkmalen der Bevölkerung in 2011 nicht möglich.

Bedingt durch Fragen, die ausschließlich aus FiD kommen, ergibt sich auch eine Notwendigkeit für Hochrechnungsfaktoren, die sich ausschließlich auf die gesamte FiD-Population (Kohorten, Screening 2010, Screening 2011) beziehen. Allerdings ist eine Kalibrierung an einen Vergleichsdatensatz wie z.B. dem Mikrozensus nicht möglich, weil es keinen Datensatz gibt, der die *Veränderungen* der einzelnen Merkmale, nach denen die FiD-Populationen ausgewählt wurden, abbilden kann. Beispielsweise kann man mit dem Mikrozensus nicht feststellen, welche Population von Frauen mit Partner im Vorjahr alleinerziehend war. Entsprechend können die angestrebten Querschnittsgewichte für die reine FiD-Population nicht nach dem für die erste Welle genutzten Verfahren gewonnen werden.

Um die gemeinsame Nutzung der FiD-Stichproben von 2010 und der FiD-Stichprobe von 2011 zu ermöglichen, ist eine Bereitstellung von FiD-spezifischen Querschnittsgewichten für die zweite Welle unverzichtbar. Hierfür wurde ein neues Verfahren entwickelt, das letztlich auf einem Ansatz beruht, der der Ausfallanalyse ähnlich ist. Den Ausgangspunkt bildet die integrierte und randangepasste Population von FiD- und SOEP-Fällen, die Merkmale der FiD-Population hat: Sie besteht aus Haushalten, die

(1) in 2010 im Niedrigeinkommensbereich waren (Screening 2010), oder
(2) in 2010 oder 2011 Mehrkindhaushalte waren (Screening 2010 oder 2011), oder
(3) in 2010 oder 2011 alleinerziehend waren (Screening 2010 oder 2011), oder
(4) in 2010 Kinder der Geburtsjahrgänge 2007-2010 hatten.

Eine so definierte Population von FiD- und SOEP-Fällen hat ein bestimmtes Bevölkerungsäquivalent, das sich mit Hilfe der gemeinsamen Hochrechnungsfaktoren errechnen lässt. Wenn nun die SOEP-Fälle aus der Population entfernt werden, müssen die Hochrechnungsfaktoren der verbleibenden FiD-Fälle angepasst werden, damit sich bei einer erneuten Hochrechnung weiterhin das Bevölkerungsäquivalent ergibt.

Allerdings ist die eben beschriebene Population nicht einfach zu bestimmen, weil es Fälle gibt, bei denen die Einordnung in die Merkmale im Jahr 2010 problematisch ist. Es handelt sich hierbei um neue Haushalte aus 2011 (sowohl FiD Screening 2011 als auch Sample J im SOEP) sowie um temporäre Ausfälle im SOEP, die 2010 nicht teilgenommen haben. Für diese Fälle muss geschätzt werden, ob sie in 2010 die oben skizzierten Merkmale besaßen oder nicht. Bei der Zugehörigkeit zur Kohorten-Stichprobe (Punkt 4) gehen wir dabei davon aus, dass die Zugehörigkeit in 2011 sich wegen der geringen zu erwartenden Veränderung auf das Jahr 2010 zurückschreiben lässt.

Dieses Verfahren lässt sich für die Merkmale (1), (2) und (3), die im Prinzip die Screening Stichproben beschreiben, nicht anwenden, weil sie im Zeitverlauf stärkeren Veränderungen unterworfen sind. Entsprechend wird hier mithilfe der Haushalte, die in beiden Jahren vorliegen, ein Logit-Modell[24] geschätzt, das anhand von 2011 erhobenen Informationen die Wahrscheinlichkeit vorhersagt, ob der Haushalt in 2010 zur Screening Stichprobe gehörte oder nicht. Die Ergebnisse dieses Models werden dann genutzt, um die Zugehörigkeit der ausschließlich in 2011 befragten Haushalte zu schätzen.

Mit dem skizzierten Verfahren lassen sich nun alle in 2011 befragten FiD- und SOEP-Haushalte in die definierten Populationsmerkmale (1)-(4) einordnen, woraus sich auch das angestrebte Bevölkerungsäquivalent ergibt. Ein Entfernen der SOEP-Fälle aus dieser Gruppe bedeutet, dass die verbleibenden FiD-Fälle im Durchschnitt höhere Hochrechnungsfaktoren erhalten müssen, damit das Bevölkerungsäquivalent sich nicht wesentlich verändert. Am einfachsten wäre hier eine Multiplikation mit einem einzigen Faktor, was aber aufgrund der unterschiedlichen Verteilung der Merkmale zwischen FiD und SOEP nicht sinnvoll ist (vgl. auch Abbildung 2a und 2b). Stattdessen wird nun mit einem Logit-Modell die Wahrscheinlichkeit berechnet[25], zur FiD-Population zu gehören, ähnlich wie in einer Ausfallanalyse die Wahrscheinlichkeit ermittelt wird, in der folgenden Welle die Studie zu verlassen. Die Kehrwert dieser Wahrscheinlichkeit ist dann der haushalts- bzw. personenspezifische Faktor, mit dem die Hochrechnungsfaktoren multipliziert und erhöht werden (entspricht der inversen Bleibewahrscheinlichkeit bei der Ausfallanalyse).

Es handelt sich bei dem hier skizzierten Verfahren um eine Neuentwicklung, die unseres Wissens bisher so noch nicht eingesetzt wurde. Dadurch, dass kein Referenzdatensatz für diese spezielle Population zur Verfügung steht, kann auch keine endgültige Evaluation der Hochrechnungsfaktoren vorgenommen werden. Die Nutzung der Hochrechnungsfaktoren kann wie sonst üblich erfolgen, wobei die hier dargestellte Problematik Berücksichtigung finden sollte. Gleichwohl weisen wir darauf hin, dass ungewichtete deskriptive Analysen mit FiD auch aufgrund der oben beschriebenen spezifischen Ziehungsdesigns keinesfalls zu empfehlen sind.

Auch für die dritte FiD-Welle (2012) werden FiD-spezifische Querschnittsgewichte bereitgestellt, die nach dem beschriebenen Prinzip aus den gemeinsamen SOEP-FiD-

[24] Kontrollvariablen dieses Modells sind die Merkmale der Screening-2010-Population in 2011, Einkommen, Anzahl der Kinder im HH, Veränderungen des Arbeitsverhältnisses zum Vorjahr und einige rückwirkend erfragte Vorjahresinformationen wie Erwerbstätigkeit, Mutterschutz, Bezug von Kindesunterhalt, Kindergeld, ALG II oder Rente.

Hochrechnungsfaktoren gewonnen werden. Die Basis bilden hier Hochrechnungsfaktoren, die ohne das im SOEP 2012 neu erhobene Sample K gebildet wurden. Auf diese Weise wird verhindert, für die Haushalte des Sample K die FiD-Populationszugehörigkeit für zwei zurückliegende Jahre schätzen zu müssen. Die Genauigkeit einer solchen Schätzung verringert sich mit zunehmendem Zeitabstand, so dass der Verzicht auf das neue Sample vorgezogen wird.

Die Version 4.0 von FiD enthält vorläufige FiD-spezifische Hochrechnungsfaktoren. Sie beruhen auf Vorwellengewichten und Bleibewahrscheinlichkeiten und berücksichtigen Anpassungen bezüglich Haushaltsaufspaltungen, temporären Ausfällen und Zuzügen von neuen Haushaltsmitgliedern. Jedoch findet keine abschließende Randanpassung statt, da das oben beschriebene Verfahren auf die SOEP-Daten der gleichen Welle angewiesen ist, die erst zu einem späteren Zeitpunkt zur Verfügung stehen.

Längsschnittgewichte

Die Erhebung der zweiten Welle markiert den Übergang in den Längsschnitt für die in 2010 gezogenen Stichproben. Um Fälle im Längsschnitt richtig hochzurechnen, müssen die Fälle, die in beiden Wellen teilnehmen in ihren Gewichten um die Fälle korrigiert werden, die nach der ersten Welle ausfallen. Diese Korrektur erfolgt über sogenannte Bleibewahrscheinlichkeiten, die sich aus den Schätzungen der Wahrscheinlichkeiten des Wiederauffindens und der erneuten Teilnahme der Haushalte ergeben. Diese Wahrscheinlichkeiten werden mittels logistischer Regression geschätzt, wobei ein Vorteil gegenüber der Berechnung der ursprünglichen Gewichte ist, dass sämtliche Haushalte der Analyse bereits im Vorjahr teilgenommen haben, und entsprechend viele eigene Daten vorliegen, die die Teilnahme in der zweiten Welle erklären können. Dennoch werden auch bei der Schätzung der Bleibewahrscheinlichkeiten Informationen von MICROM genutzt. Der Kehrwert der geschätzten Teilnahmewahrscheinlichkeit bildet dann das Bleibegewicht, das als Korrektur auf das ursprüngliche Gewicht multipliziert wird, um Analysen im Längsschnitt hochrechnen zu können.

Nutzung der Hochrechnungsfaktoren

Das Hochrechnungskonzept von FiD erlaubt es dem Nutzer, die FiD-Stichprobe separat zu analysieren oder sie in Kombination mit den Daten des SOEP zu verwenden und gemeinsam

[25] Die Variablen dieses Modells bilden die Zusammensetzung der Haushalte bezüglich der FiD-Merkmale in den Jahren 2010 und 2011 ab und werden ergänzt durch die Wohnregion (Ost / West) sowie im Fall der Personengewichte um Altersgruppen.

hochzurechnen. Insgesamt sind in der FiD-Datenweitergabe vier verschiedene wellenübergreifende Datensätze mit Hochrechnungsfaktoren enthalten. Diese sind:

1. ***hhrf***: Hochrechnungsfaktoren innerhalb der FiD-Daten auf Haushaltsebene
2. ***phrf***: Hochrechnungsfaktoren innerhalb der FiD-Daten auf Personenebene
3. ***hhrf_fidsoep***: integrierte Hochrechnungsfaktoren für die Kombination von SOEP- und FiD-Daten auf Haushaltsebene
4. ***phrf_fidsoep***: integrierte Hochrechnungsfaktoren für die Kombination von SOEP- und FiD-Daten auf Personenebene

Die Datensätze ***hhrf*** und ***phrf*** enthalten jeweils die Identifikatoren (HHNR und HHNRAKT bzw. PERSNR), die es erlauben eine Beobachtung zu identifizieren und die Gewichte anderen Daten zuzuspielen. Die folgenden Gewichte sind hier enthalten:

hhrf

f$$hhrf*	HRF für die gesamte FiD-Stichprobe 20$$ auf Haushaltsebene
f11hbleib	Inverse Bleibewahrscheinlichkeit für 2011 auf Haushaltsebene
f12hbleib	Inverse Bleibewahrscheinlichkeit für 2012 auf Haushaltsebene
f13hbleib	Inverse Bleibewahrscheinlichkeit für 2013 auf Haushaltsebene
f10hhrf_sc10	HRF für die Screening-Stichprobe 2010 auf Haushaltsebene (nur 2010)
f10hhrf_co10	HRF für die Kohorten-Stichprobe 2010 auf Haushaltsebene (nur 2010)
f11hhrf_sc11	HRF für die Screening-Stichprobe 2011 auf Haushaltsebene (nur 2011)

phrf

f$$phrf*	HRF für die gesamte FiD-Stichprobe 20$$ auf Personenebene
f11pbleib	Inverse Bleibewahrscheinlichkeit für 2011 auf Personenebene
f12pbleib	Inverse Bleibewahrscheinlichkeit für 2012 auf Personenebene
f13pbleib	Inverse Bleibewahrscheinlichkeit für 2013 auf Personenebene
f10phrf_sc10	HRF für die Screening-Stichprobe 2010 auf Personenebene (nur 2010)
f10phrf_co10	HRF für die Kohorten-Stichprobe 2010 auf Personenebene (nur 2010)
f11phrf_sc11	HRF für die Screening-Stichprobe 2011 auf Personenebene (nur 2011)

hhrf_fidsoep

f$$hhrf_soep	HRF SOEP integriert mit FiD 20$$ auf Haushaltsebene
f11hbleib	Inverse Bleibewahrscheinlichkeit für 2011
f12hbleib	Inverse Bleibewahrscheinlichkeit für 2012

phrf_fidsoep

f$$phrf_soep	HRF SOEP integriert mit FiD 20$$ auf Personenebene
f11pbleib	Inverse Bleibewahrscheinlichkeit für 2011
f12pbleib	Inverse Bleibewahrscheinlichkeit für 2012

* In Version 4.0 **vorläufige** Hochrechnungsfaktoren für 2013

Anhang: Neue Gewichte 2010-2012 ab der Weitergabe FiDv3.1

Folgende Punkte führen zu Veränderungen in den Gewichten ab der Weitergabeversion 3.1

1) Änderung im Umgang mit temporären Ausfällen im SOEP und in FiD
2) Neue Gewichte für die Kohorten-Stichprobe der im 1. Quartal 2010 geborenen Kinder

1. Temporäre Ausfälle

Sowohl im SOEP als auch in FiD gibt es sogenannte temporäre Ausfälle von Haushalten. Dies sind Haushalte, die in einem Jahr an der Befragung nicht teilnehmen, in einem folgenden Jahr erneut kontaktiert werden und dann an der Befragung teilnehmen. Für einen Haushalt, der genau eine Welle „aussetzt", wird in der Welle des Wiedereintritts (t) der Hochrechnungsfaktor bei der letzten Teilnahme (t-2) mit der inversen Bleibewahrscheinlichkeit der Welle des Wiedereintritts (t) multipliziert. Diese Bleibewahrscheinlichkeit basiert auf der Bruttostichprobe von Welle t, die auch die temporären Ausfälle beinhaltet. Ein Problem ergibt sich nun für diese Haushalte in der Gewichtung für die Welle des Wiedereintritts, wenn während des temporären Ausfalls (t-1) eine neue Stichprobe in das Gesamtsample integriert wurde (beispielsweise Sample J in 2011 für das SOEP oder die Screening Stichprobe 2011 in FiD). Sofern eine neue Stichprobe auch Haushalte mit denselben Charakteristika der bereits vorhandenen Stichprobe enthält, müssen bei der Integration die Hochrechnungsfaktoren der „alten" Haushalte reduziert werden, um zu berücksichtigen, dass die gleiche Grundgesamtheit nun von einer größeren Anzahl an Fällen in der Stichprobe repräsentiert wird. Diese Reduzierung wurde bisher nicht für die temporären Ausfälle vorgenommen, weil diese in der jeweiligen Welle nicht im Sample enthalten waren. In der jetzigen Weitergabe wurde diese Ungenauigkeit behoben, so dass auch bei temporären Ausfällen die Reduzierung der Hochrechnungsfaktoren, die durch die Integration neuer Stichproben im Jahr des Ausfalls nötig wird, vorgenommen wird. Für FiD betrifft diese Änderung 181 Fälle, die in 2010 und 2012 teilnehmen, und 2011 ausgesetzt haben.

2. Neue Gewichte für die Kohorten-Stichprobe der im 1. Quartal 2010 geborenen Kinder

Die Kohorten-Stichprobe der 2010 geborenen Kinder besteht aus Haushalten, die 2010 aus Einwohnermeldedaten gezogen wurden und mindestens ein Kind beinhalten, das zwischen Januar und März 2010 geboren wurde. Die Hochrechnung (siehe auch Dokumentation „hhrf_phrf" der Version FiDv3.0, S.15) erfolgte bisher unter der Maßgabe, die vorliegenden Haushalte auf die gesamte Geburtskohorte hochzurechnen und damit die Kinder der ersten drei Monate als repräsentativ für das gesamte Jahr zu betrachten. Berechnungen der FiD-

Arbeitsgruppe am SOEP haben jedoch eine Revision nahegelegt: dabei werden die Haushalte der Kohorte 2010 entsprechend ihrem realen Anteil gewichtet und deren Hochrechnungsfaktoren im Schnitt auf ca. ein Viertel des vorherigen Wertes sinken. Dadurch kann es bei der Betrachtung bestimmter Variablen zu Unterschieden gegenüber der bisherigen Gewichtung kommen, die sich auf die hochgerechnete Anzahl an Kindern bzw. an Haushalten mit Kindern im ersten Geburtsquartal 2010 zurückführen lassen.

Dieser Sachverhalt wird im Folgenden an einem Beispiel verdeutlicht. Haushalte aus FiD und SOEP, die 2010 befragt wurden und in denen Kinder im Alter von 5-12 Monaten leben, kommen aus allen Stichproben. Der Anteil der Kohorten-Stichproben mit 2010 geborenen Kindern liegt dabei bei knapp 49%. Bei dieser Untergruppe handelt es sich um Haushalte mit jungen Kindern, deren Eltern zwischen Juni und September 2010 befragt wurden, wobei die Kinder bei der Befragung meistens zwischen fünf und acht Monaten alt waren. Betrachtet man nun die durchschnittliche Arbeitsmarktpartizipation von Frauen in Haushalten mit Kindern im Alter von 5-12 Monaten spielt dieses Alter eine Rolle: mit den korrigierten Gewichten steigt die Arbeitsmarktpartizipation von Frauen von 13,9% auf 16,6%, da die hochgerechnete Masse der Frauen mit jungen Kindern (die aus der Kohorte 2010) relativ zu denen mit älteren Kindern (Kohorte 2009 und andere) durch die Korrektur kleiner wird. Gegenläufige Effekte treten beispielsweise in 2011 bei der Betrachtung von Haushalten auf, deren Kinder zwischen 13 und 18 Monaten alt sind: hier geht der Anteil der arbeitenden Frauen von 42,8% (alte Gewichte) auf 37,0% (neue Gewichte) zurück. Dies lässt sich mit der Verteilung der Kinder in dieser Altersgruppe begründen: von den knapp 2,300 Kindern in dieser Gruppe kommen 60% aus der Geburtskohorte 2010. Innerhalb dieser Gruppe sind 99% zwischen 15 und 18 Monaten alt, ein Alter, in dem Mütter nach der Elternzeit wieder anfangen zu arbeiten. Durch die neue Gewichtung werden die Mütter der jüngeren Kinder relativ stärker berücksichtigt, so dass sich eine leichte Reduzierung der durchschnittlichen Arbeitsmarktpartizipation ergibt.

Veränderungen lassen sich auch für andere Variablen erwarten, in denen bestimmte Charakteristika der Kohorten-Stichprobe der im ersten Quartal 2010 geborenen Kinder, insbesondere zusammenhängend mit dem Alter dieser Kinder, eine Rolle spielen. Die bisherigen Tests und Vergleiche legen allerdings nahe, dass sich alle Änderungen innerhalb des durch die Konfidenzbänder vorgegebenen Rahmens bewegen und somit keine substantiellen Änderungen an bisherigen Ergebnissen zu erwarten sind.

Documentation *mipinc* and *mihinc*

Multiple imputation of income variables

Mathis Fräßdorf (geb. Schröder)

Contents

Introduction

FiD employs multiple multivariate imputations for income and other monetary variables in the data. This is on two levels – the household and the individual level. For multiple imputations, it is important to jointly impute all variables, to allow for cross-variable dependencies. Hence, within each level, all variables of interest are imputed at the same time. Table 1 shows the variables which are imputed and included in the datasets ***$mipinc*** and ***$mihinc*** over data collections in 2010-2013, along with the respective number of values that had to be imputed in comparison to the non-missing values.[26]

Table 1a: Imputed variables on the individual level (dataset ***$mipinc***)

		Ratio of missings				imputed values			
Name	**Description**	**2010**	**2011**	**2012**	**2013**	**2010**	**2011**	**2012**	**2013**
$PNETINC	Net labour income	0.105	0.054	0.046	0.037	520	288	239	187
$PGROINC	Gross labour income	0.192	0.069	0.06	0.056	947	370	309	286
$PMATBEN	Maternity benefits	0.028	0.014	0.015	0.000	27	4	3	0
$PALIMON	Alimony payments	0.036	0.013	0.012	0.008	31	12	10	6
$PPENS	Pensions/Retirement benefits	0.063	0.037	0.042	0.064	7	5	6	10
$PUEBEN	Unemployment benefits (ALG1)	0.043	0.036	0.027	0.026	7	4	3	3
$PWIDOW	Widow(er) pensions	0.039	0.060	0.068	0.036	3	5	6	3

Note: Non-responding household members (PUNR) are not considered in the table, but values are imputed. Given that large parts of the data are missing, these imputations should be handled with care. Also note that "imputed values" may vary by implicate, as the receipt of an income group is also subject to missing values and hence imputation.

Source: FiD 2014. No further distribution without the explicit consent of the author.

Table 1 shows quite substantial differences in the size of the missing value problem depending on the variable of interest as well on the year of the survey. In general, the missings rates decrease with increasing years and thus fewer values need to be imputed, which is not unusual for the later waves of a panel survey. All in all, the values are largely comparable to the missing ratios in the SOEP. In the individual data, the variable $PGROINC has the largest amount of missing values in each year. Part of this is due to the calculation procedure for this variable. Instead of only imputing the gross income, we also restrict it to be equal or larger than the net income. This leads to an imputed value, whenever $PGROINC *or*

[26] Note that we use the complete FiD distribution for the imputations and the descriptions as shown here, i.e. including those cases with a weight of zero. Users should be careful to be aware of these cases when using the data. Note also that the variable f10plsyrinc, "last year's gross income", is no longer included in the imputations, as users should prefer to use the stated by the respondent or its imputed value.

$PNETINC are missing. Such a procedure is unique for $PGROINC, and leads to an increase in the missing percentage. On the other hand, the missing ratios for maternity benefits and child support are very low, as are the total numbers which had to be imputed for these variables. This is important for the general evaluation purpose of FiD, as reported data are usually preferred to imputed data.

Table 1b: Imputed variables on the household level (dataset ***$mihinc***)

		Ratio of missings				**imputed values**			
Name	**Description**	**2010**	**2011**	**2012**	**2013**	**2010**	**2011**	**2012**	**2013**
$HHINC	Monthly net household income	0.053	0.025	0.023	0.023	243	114	96	89
$HCHDBEN	Monthly child benefits ('Kindergeld')	0.010	0.002	0.001	0.001	46	9	3	5
$HCHDADD	Monthly added child benefits ('Kinderzuschlag')	0.232	0.071	0.03	0.034	54	11	4	4
$HUEBEN2	Monthly long-term UE benefits ('ALG2')	0.034	0.011	0.012	0.013	31	9	8	8
$HHOSBEN	Monthly housing benefits ('Wohngeld')	0.057	0.015	0.009	0.015	23	7	3	4
$HCARBEN	Monthly care benefits (‘Pflegegeld)	0.085	0.047	0.02	0.031	5	2	1	2
$HHELBEN	Monthly other benefits (‘Hilfe Lebenslagen’)	0.182	0.029	0.019	0.085	16	2	1	5
$HAGETRN	Monthly age benefits (‘Grundsicherung Alter’)	0.043	0.087	0.000	0.056	1	2	0	1
$HRENT	Monthly rent payments	0.006	0.002	0.003	0.001	16	6	7	3
$HUTIL	Monthly utility payments	0.597	0.580	0.582	0.56	1709	1596	1457	1300
$HHEAT	Monthly heating payments	0.221	0.176	0.184	0.183	634	483	461	424
$HELECTR	Monthly electricity payments	0.090	0.071	0.064	0.057	258	196	159	132
$HCREDIT	Monthly credit payments	0.041	0.015	0.021	0.006	59	22	26	8

Note: “imputed values” may vary by implicate, as the receipt of an income group is also subject to missing values and hence imputation.

Source: FiD 2014. No further distribution without the explicit consent of the author.

On the household level, first note that $HRENT, $HUTIL, $HHEAT and $HELEC are imputed only for those renting the accommodation they live in. The last three variables for renters have relatively large missing percentages in all years, with the highest on the utility payments, i.e. payments for trash removal, water, etc. One reason for this high percentage is

that - due to questionnaire routing - utilities have to be imputed whenever the household head states that rent does not include utilities. However, other variables have low missing percentages – notably household income, with 5.3% missing in 2010 and 2.5% or less missing since 2011.

Why then use multiple imputations? Missing information itself is a problem in any analysis if it is not missing at random, i.e. if the sample that is left by the missing data generation process is selective or different from the sample that contains missing data. Imputations per se are meant to deal with this problem, where the usual assumption is, that data from the non-missing population can be used to infer the values for the missing observations (see e.g. Starick and Watson, 2007, or Little, 1992). However, single imputations are problematic in certain aspects – while the predictions are usually good, the standard errors in analyses are reduced (see e.g., Rubin, 1996, or Schafer and Graham, 2002). This is a feature of "pretended" certainty as only one value per missing observation is imputed. Multiple imputations on the other hand try to mitigate this issue by not only imputing one, but multiple values (so-called implicates) for each missing observation. This leads to better variance estimates, as the uncertainty of the data is kept. Even though the computational effort is somewhat increased – both for the data provider and for the researcher using the imputations later – we believe that the benefits outweigh these costs.

The rest of the documentation describes the procedures applied for the multiple imputations on both levels. We then provide some measures which we used to evaluate the performance of our imputations. Finally, we briefly describe how to use the multiple imputations in practice.

Methods

The main procedure for the imputations is identical for the household and the individual level:

(1) The first step is to generate a dataset, which contains all necessary information for the imputations.

(2) Then, as we are using a method of chained equations (see below), for each variable that is to be imputed, a parsimonious model needs to be specified.

(3) The third step then is the main part – imputing all variables according to the models in step 2.

(4) The imputations are evaluated in the fourth step.

(5) Finally, the data are prepared for their distribution to the user.

In step 1 the basic datasets are used to create variables which could be used in the imputation procedure. Data are taken from ***$h***, ***$p***, ***$pbrutto***, ***$hbrutto***, and ***$lela***. All variables were taken from the original questionnaires, although some of them are then restructured similarly to the generating process. For all income variables natural logarithms are used instead of the actual values because of the skewed distributions. Some household level characteristics are added to the imputations on the individual level. These include size of the household, number of children, whether the household belongs to the Cohort-Sample as well as the cross-sectional weights. On the household level, individual variables are used as well: these included whether any household member receives certain incomes, is working, as well as the partner structure in the household.[27] Table 2 provides some more detail on the variables included on each level.

Step 2 then involves setting up equations for each variable with missing values. Here the focus is on the predictive power of each equation, not on substantial modelling. Hence, the imputations do not consider endogeneity or simultaneity issues, but are solely interested in better predictions, as long as the convergence criteria are met. For example, we tested whether the incomes could be used for prediction, i.e. use net income to predict gross income and vice versa. However, convergence was a problem in these models, likely because of the large overlap of missing values in both incomes on the one hand and the high predictive power of the variables for each other on the other hand. The predictive power did not suffer too much by not including them in the respective regressions. Starting with 2011, it is possible to use information from other years in the imputation. In principle, there are no prior reasons not to

[27] Until FiD v3.1, the household imputations were independent of the individual imputations. This meant that individual variables, which were aggregated on the household level, had to be imputed in case one person's information on that variable was missing. To reduce the computational effort, we use the results from the individual imputations (i.e. the first implicate) for household aggregated values starting with FiD v4.0.

take all available years into account in any year's imputation. However, we only consider the previous year in the imputations (i.e. 2011 uses 2010, 2012 uses 2011, 2013 uses 2012), because the computational effort proved to be too large when considering multiple years.

The selection of equations is conducted with the *stepwise* command in Stata™, again keeping in mind that good prediction is the key. In specifying the *stepwise* command, variables with a significance level above 0.10 are removed from the model. This leads to a considerable reduction in complexity for most variables that needed imputation. With these specified models, the imputation for all variables is executed with the *mi impute chained* command.

We use Stata™ (release 13) for the imputations in step 3, and here the *mi* package first introduced with Stata 12. The imputations are done in a *Multivariate Imputation by Chained Equations* (or *mice*) procedure (see also van Buuren et al., 2006 for a detailed description). The basic idea of the m*ice* procedure is to use a regression equation for variable *x* to predict missing values in *x*, use these predicted values for the next equation when modelling variable *y*. The predicted values for the missing observations are then used in the next equation, until all equations have been estimated and one "cycle" is completed. The next cycle starts with the same sequence, but uses the predicted values from the first cycle. These cycles or iterations are repeated, until some measure of convergence is achieved. Also see the Stata (2013) for details.

Table 2a: Variables used in the individual-level imputations

Context	**Variables included**
Income	Natural logarithms of: net labour income; gross labour income; current maternity benefits; current child support
Job characteristics	Hours worked; overtime; trained for job; job satisfaction; working full or part-time; being in vocational training; tenure; firm size; occupation
Demographics	Age; gender; foreigner; east/west; education; health; marital status
Household characteristics	Age, gender, and foreign status of household head; number of children; household size; number of non-respondents in household; cohort dummy; household weight, federal state; size of community (BIK)
Satisfaction	Current satisfaction with different aspects in the respondent's live

The imputations themselves are restricted to those cases where imputation is useful. For example, labour income is not imputed for individuals who do not work, and similarly, rent is not imputed for owners. Note that if there are missing values in any of the independent

variables, imputations are imperative for these variables as well. For example, education has to be imputed for some cases. We do not distribute these variables, as they usually concern only a few cases and including them would increase datasets by an unnecessary extent.

The imputations for the individual income variables are done with overall 750 cycles and five implicates, using nearest neighbour matching with a matchpool of 3 observations to impute the data. While this takes considerable time, the evaluation shown below provides evidence that this large amount of cycles is quite useful to reach convergence. On the household level, we use 750 cycles, while also providing five implicates for each imputed variable.

Table 2b: Variables used in the household-level imputations

Context	**Variables included**
Income	Natural logarithms of: household income; monthly child benefits; added child benefits; long term unemployment benefits; housing benefits; rent; utility payments; heating payments; electricity payments; credit payments. Additionally the categories of household income from the unfolding bracket sequence.
Expenditures	Natural logarithms of: rent; utility payments; heating payments; electricity payments; credit payments; payments on durable and non-durable goods
Individual incomes	Natural logarithms of all income sources.
Individual aggregates	Any employed; working status; highest education; any married, divorced, or widowed;
Household characteristics	Age, gender, and foreign status of household head; number of children; household size; number of non-respondents in household; cohort dummy; household weight; number of rooms; square meters of housing unit; federal state; size of community (BIK)

The *mi chained* procedure itself does not check convergence or goodness of fit of the imputed variables, so this is up to the researcher. We use two main criteria in step 4 to evaluate the performance of our imputed values. First, we implemented the *Gelman-Rubin-Brooks* criterion (see Brooks and Gelman, 1998) in Stata™ to test convergence, which compares the within-variance within a chain of imputations with the between variance across the chains (see the next section for details). We then compared the distribution of imputed variables to the non-missing observations. Some caution needs to be taken here. By comparing the distribution of non-imputed values with those which are imputed, one should not expect them to be equal, as individuals with missing values may be different than those without missing

values. Hence a comparison of distributions may only serve as a rough eyeball test for whether there are significant problems or not. The following section provides some details on the evaluation as well as the results.

Finally, in step 5, the data are prepared for the distribution. This involves generating the imputation flags which indicate which values are imputed. One variable, $PGROINC has to be specifically constructed from the imputations of other variables, as mentioned above. Some variables are integrated into ***$pgen*** and ***$hgen***: these are $PNETINC and $PGROINC which are renamed to LABNET$$ and LABGRO$$ in ***$pgen***. $HRENT, $HUTIL, $HHEAT and $HELEC become part of ***$hgen***. In addition, the variables ROOM$$ and SIZE$$ are taken from the imputations and added to ***$hgen*** as well.[28] For each of the previous variables, the first and only the first implicate is used, even though this counteracts the initial purpose of the multiple imputations. We do not recommend using only one implicate, but recognize that some users may find it easier to deal with only one implicates instead of five. Also, we follow the SOEP logic with this step. As in the SOEP, for $HHINC five implicates are included for each household as I1HINC to I5HINC in ***$hgen***. Variables included in the two multiple imputation datasets (***f10mipinc*** and ***f10mihinc***) are listed and described below.

Evaluation of imputations

The imputations for each variable are evaluated by criteria concerning the convergence of the different iterations on the on hand and the final distribution of the imputed values on the other.

For the evaluation of convergence we use the criteria as specified in Brooks and Gelman, 1998. The authors identify the following measures, for which we specify *m* as the number of implicates, in our case *m*=5. *n* is the number of cycles or iterations the imputations have run until convergence is reached; in the individual case, *n*=7500, while in the household case, *n*=750. Then the between-implicate variance *B/n* is defined as follows (see Brooks and Gelman, page 436):

$$(1) \qquad B/n = \frac{1}{m-1}\sum_{j=1}^{m}\left(\overline{\varphi}_{j.} - \overline{\varphi}_{..}\right)^2$$

and the within-implicate variance, *W*, is

$$(2) \qquad W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}\left(\overline{\varphi}_{jt} - \overline{\varphi}_{j.}\right)^2 .$$

Brooks and Gelman then continue to define the posterior variance estimate:

[28] Note that in version 4.0 of the FiD data, no imputations were necessary for the number of rooms in either year.

(3) $$\hat{V} = \frac{n-1}{n}W + \frac{m+1}{mn}B$$

Finally, the potential scale reduction factor, or PSRF, is given as the Ratio of V and W:

(4) $$\hat{R} = \frac{\hat{V}}{W} = \frac{n-1}{n}W + \frac{m+1}{mn}B\frac{1}{W} = \frac{n-1}{n} + \frac{m+1}{nm}\frac{B}{W}$$

(Note that Brooks and Gelman apply a correction factor *d* to their statistic, which for simplicity is ignored here.)

Brooks and Gelman define convergence to be achieved if:

a) $\hat{V}$ stabilizes as a function of *n*,

b) *W* stabilizes as a function of *n*, and

c) $\hat{R}$ approaches 1, implying that $\hat{V}$ and *W* converge to one another .

The convergence is monitored graphically, where the number of iterations is plotted against the three statistics above. (Note that we use the same approach as Brooks and Gelman in their paper, where we only consider certain steps in the convergence process, and hence restrict the number of points to 40 at all times.) $\hat{R}$, $\hat{V}$, and *W* can be computed for different statistics of the respective variable. We consider the two most common here, the mean and the standard deviation.

To determine the performance of our final imputations, we compare the distribution of the five implicates with the original, non-imputed distribution. Note, however, that an argument can be made for the distributions not to match, as the population with missing observations may be different in the variable(s) of interest to the population with non-missing values. E.g., if people with high incomes are more likely not to report their income, the distribution for the imputed values should be shifted right compared to the non-imputed observations.

The graphs for the Gelman-Rubin-Brooks criterion for mean and standard deviation as well as the comparison of the densities can be found with the variable descriptions below for the most important variables: net and gross individual income as well as household net income. Graphs for other variables are available on request.

Using multiple imputations

The datasets of multiply imputed variables are distributed separately from all others. They only contain the imputed variables and their imputation flags, indicating whether a value had been imputed or not. Individual data is identified via the PERSNR identifier, while households (in both datasets) are identified through $HHNR. To use the imputation datasets for analysis, they need to be merged with other data using the respective identifiers.

As multiple imputations become more common in social science datasets, statistical packages now allow for their use. When using multiple imputations, means, coefficients and variances have to be adjusted to take into account the different implicates. Basically, any analysis has to be conducted as many times as there are implicates, and then coefficients can be combined and joint variances calculated. Following Rubin (1987), the coefficient of *m* regressions using each of the implicates *j=1,…,m* lead to a combined estimate of

$$\bar{\beta}_m = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j$$

For the variance, the within-imputation variance needs to be combined with the between-imputation variance. The former is derived as the mean of the variances from each imputation regression

$$\bar{W}_m = \frac{1}{m}\sum_{j=1}^{m} Var\left(\hat{\beta}_j\right),$$

while the latter is computed as the variance of the coefficients over the *m* implicates:

$$B_m = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_j - \bar{\beta}_m\right)$$

The joint variance is then given by

$$T_m = \bar{W}_m + \frac{m+1}{m}B_m$$

While the user may calculate these statistics him or herself, statistical software provides these and more applications for multiple imputations. Since the distribution of Stata 11 a complete package on multiple imputations is included – allowing to impute as well as using the imputations in analyses. See the *mi* command in that package. In addition, there are several added programs for analyzing multiple imputed data, for example *mim*. Within SAS, PROC MIANALYZE combines the results of analyses on the different data sets. IVEware is a set of routines that can be launched from SAS or run independently using data from many sources. You can use the IVEware module *regress* to perform multiple imputation analysis.

Variables in $mipinc

_MI

Variable label	**"observation number"**
Variable format	5-digit integer
Comment	_mi identifies observations belonging to the same household. For each person, there are 6 observations – one original and 5 imputed values.

_MJ

Variable label	**"imputation number"**
Variable format	4-digit integer
Comment	_MJ identifies the implicates for each person and variable. The range is from 0 to 5, where 0 denotes the original observation and 1-5 identify the respective imputations.

PUNR$

Variable label	**"Non-responding household member"**
Value label	PUNR$ (0) Regular respondent (1) Non-respondent
Variable format	1-digit integer
Comment	This variable identifies, whether an observation stems from a non-responding household member (partial unit non-response, PUNR) or not. Note that values are imputed for all household members, even if they decide not to participate in the survey.

$PNETINC

Variable label	**"monthly net labor earnings (imputed)"**
Variable format	5-digit integer
Comment	The original question in the person questionnaire is about the net income from labor earnings for each working individual. Hence "-2" (does not apply) is set for all non-working individuals. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$PGROINC

Variable label	**"monthly gross labor earnings (imputed)"**
Variable format	5-digit integer
Comment	The original question in the person questionnaire is about the gross income from labor earnings for each working individual. Hence "-2" (does not apply) is set for all non-working individuals. The imputation

was conducted using the difference between net and gross income, imputing it, and then adding it to the imputed net income. As mentioned above, this leads to more imputed values than were originally missing. However, the consistency here seemed more important.

$PMATBEN

Variable label **"maternity benefits (imputed)"**
Variable format 4-digit integer

Comment The origin of this variable is the question for the maternity benefit while being on some sort of maternity leave ("Elterngeld", "Erziehungsgeld"). "-2" (does not apply) is set for all individuals who state not to receive any maternity benefits. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$PALIMON

Variable label **"child alimony (imputed)"**
Variable format 4-digit integer

Comment This variable originates in the question for child support received individually ("Kindesunterhalt"). Note that "-2" (does not apply) is set for all individuals who do not receive any child support. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$PUEBEN

Variable label **"unemployment benefits(imputed)"**
Variable format 4-digit integer

Comment This variable originates in the question for unemployment benefits ("Arbeitslosengeld"). Note that this is not "ALG2", which is received on the household level. Note that "-2" (does not apply) is set for all individuals who do not receive any benefits. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$PWIDOW

Variable label **"widow(er) pensions (imputed)"**
Variable format 4-digit integer

Comment This variable originates in the question for pensions received as a widow or orphan ("Witwen/Waisenrente"). Note that "-2" (does not apply) is set for all individuals who do not receive any pensions. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$PPENS

Variable label	**"pensions/retirement (imputed)"**
Variable format	4-digit integer
Comment	This variable originates in the question for pensions and retirement benefits received individually ("Rente/Pension"). Note that "-2" (does not apply) is set for all individuals who do not receive any pensions. The natural logarithm of this variable was imputed, and then converted back to the monetary values.

I_$PNETIC

Variable label	**"f$pnetic is imputed"**
Value label	I_$PNETIC (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all non-working individuals.

I_F10PGROINC

Variable label	**"f$pgroinc is imputed"**
Value label	I_$PGROINC (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all non-working individuals.

I_$PMATBEN

Variable label	**"$pmatben is imputed"**
Value label	I_$PMATBEN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all individuals who do not receive any maternity benefits.

I_$PALIMON

Variable label	**"$palimon is imputed"**
Value label	I_$PALIMON (0) Observed value (1) Imputed value
Variable format	1-digit integer

Comment	Note that "-2" (does not apply) is set for all individuals who do not receive any child support.

I_$PUEBEN

Variable label	**"$pueben is imputed"**
Value label	I_$PUEBEN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all individuals who do not receive any unemployment benefits.

I_$PWIDOW

Variable label	**"$pwidow is imputed"**
Value label	I_$PWIDOW (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all individuals who do not receive any widow(er) benefits.

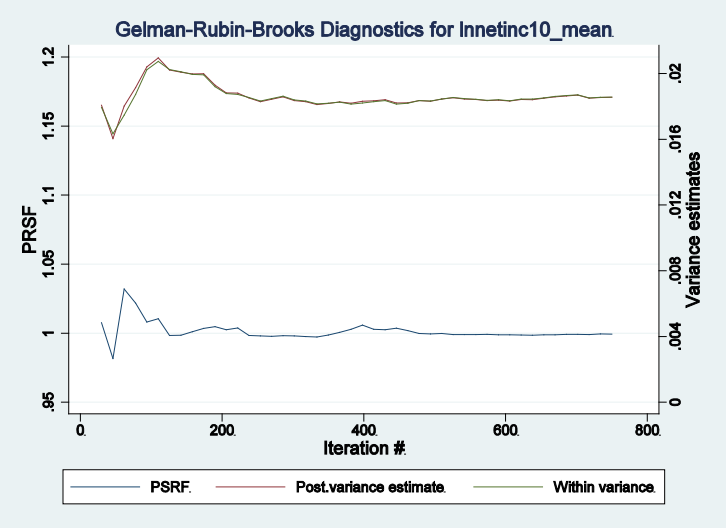
I_$PPENS

Variable label	**"$ppens is imputed"**
Value label	I_$PPENS (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all individuals who do not receive any pension benefits.

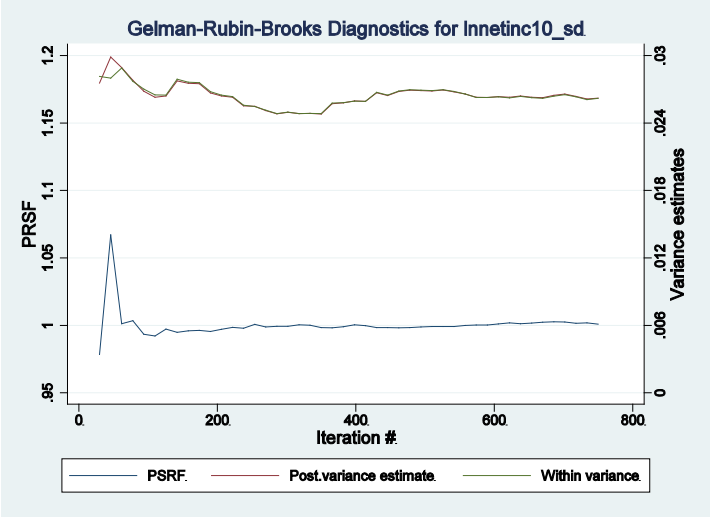
Evaluation graphs for $mipinc

F10PNETINC

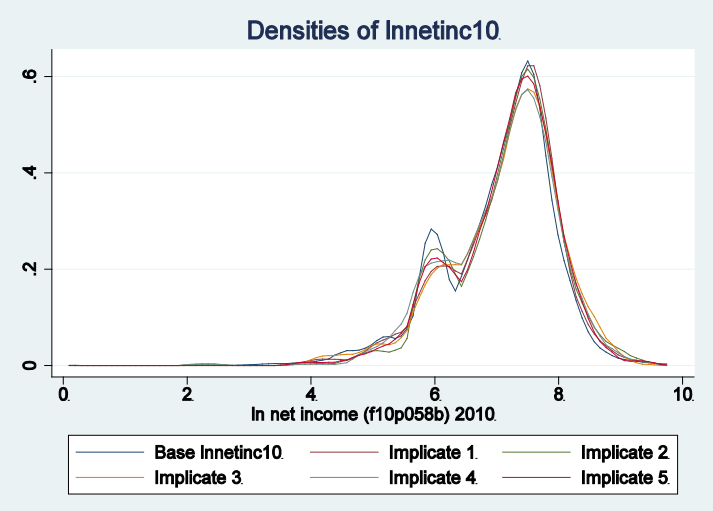
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

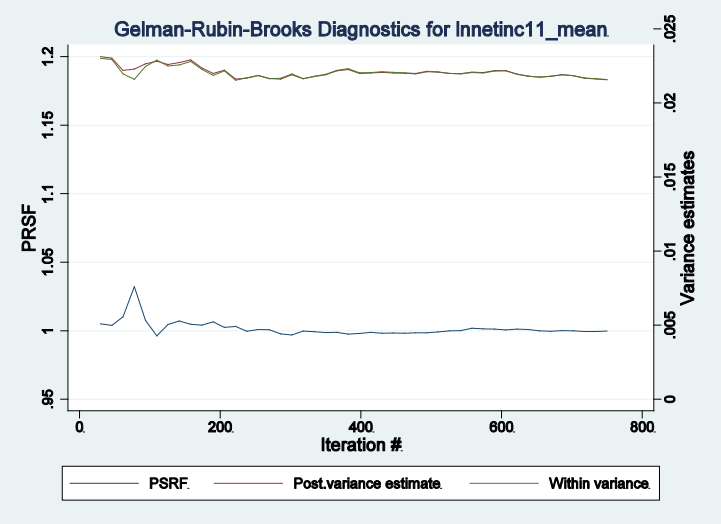


Comparison of Densities between original distribution (blue) and the five implicate

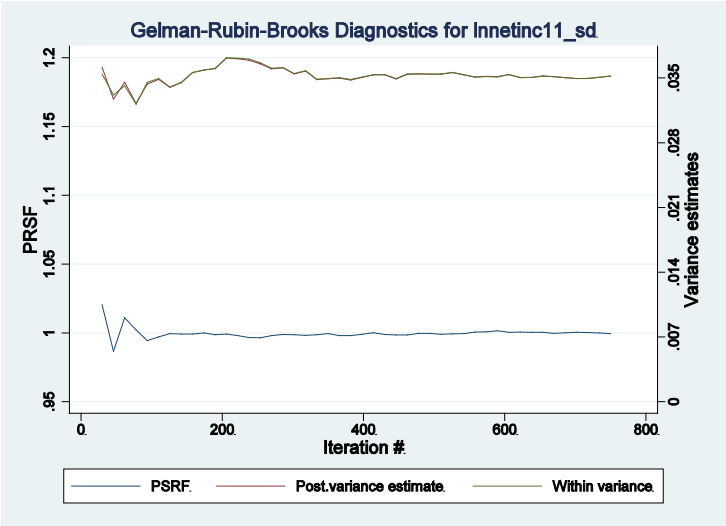


F11PNETINC

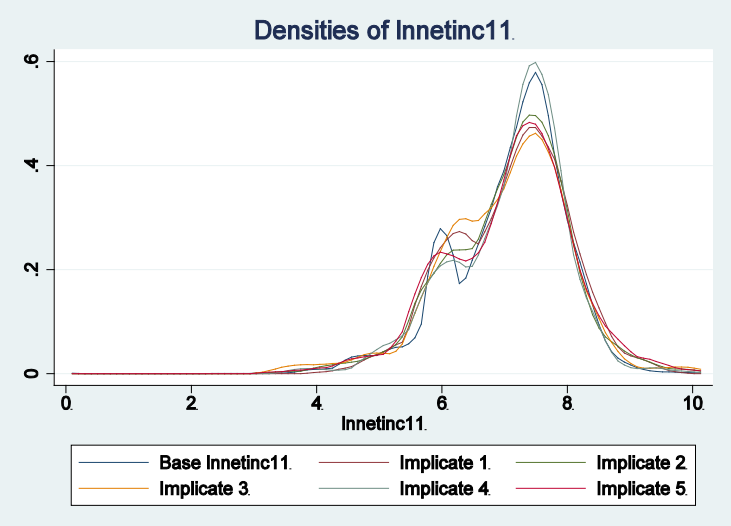
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

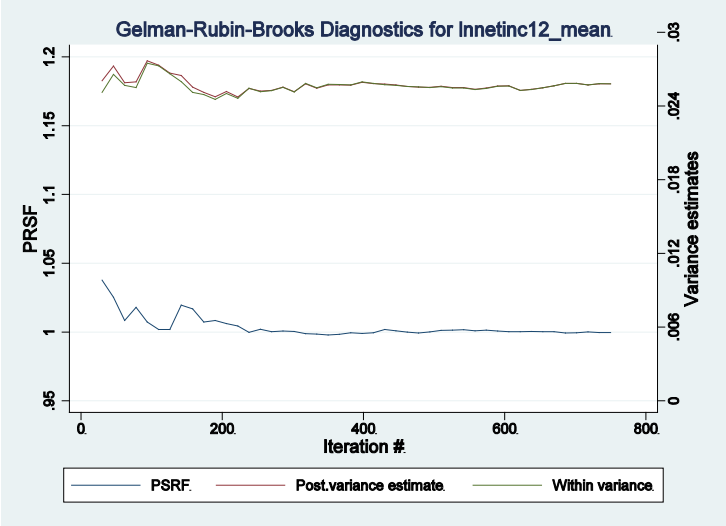


Comparison of Densities between original distribution (blue) and the five implicates

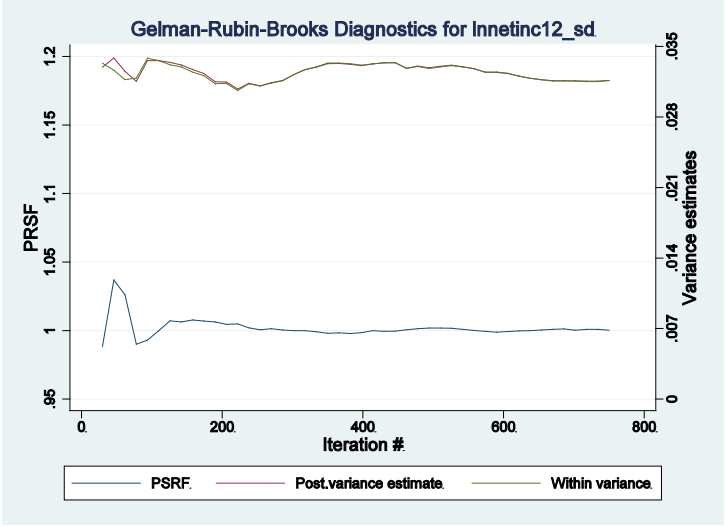


F12PNETINC

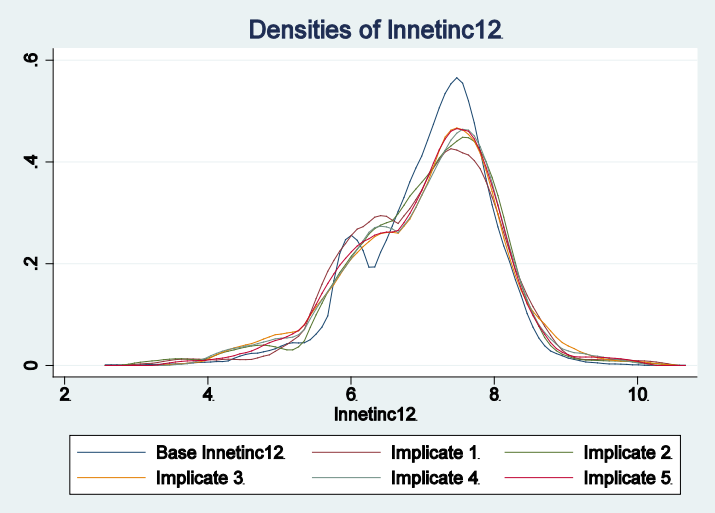
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

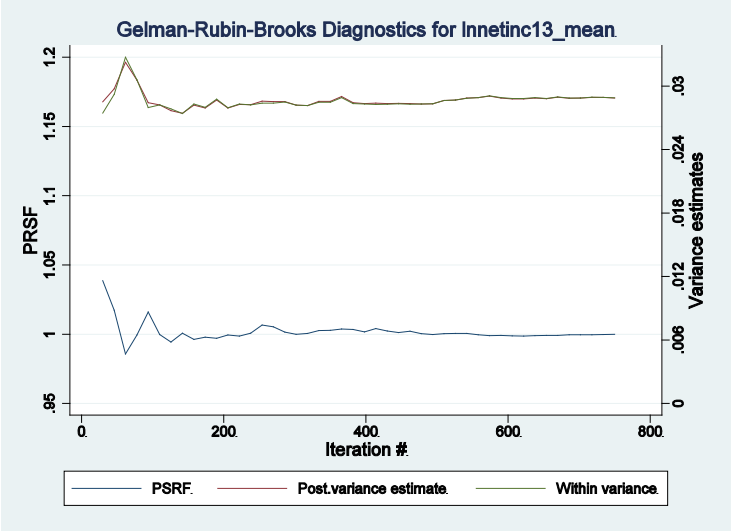


Comparison of Densities between original distribution (blue) and the five implicates

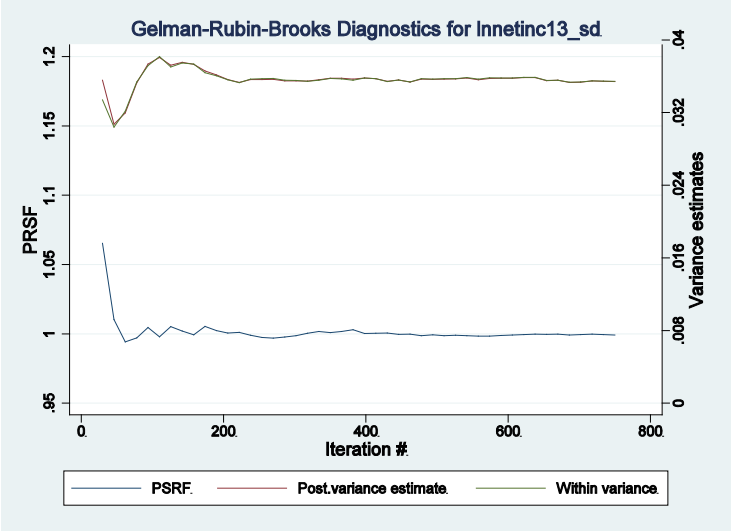


F13PNETINC

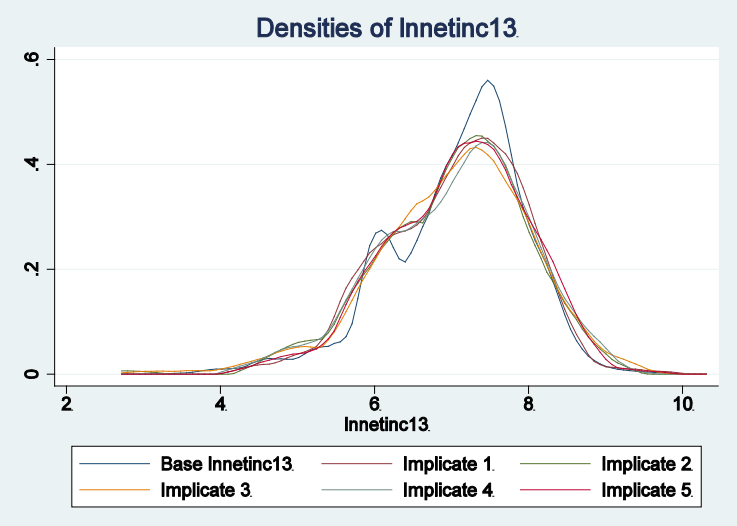
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

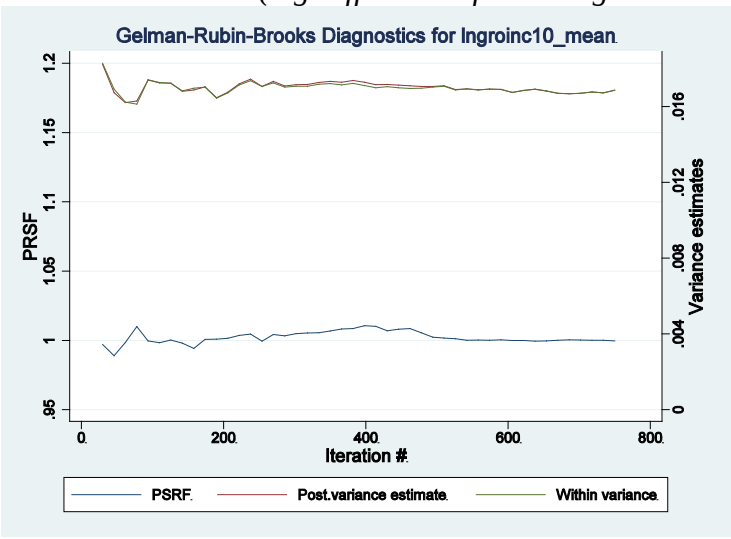


Comparison of Densities between original distribution (blue) and the five implicates

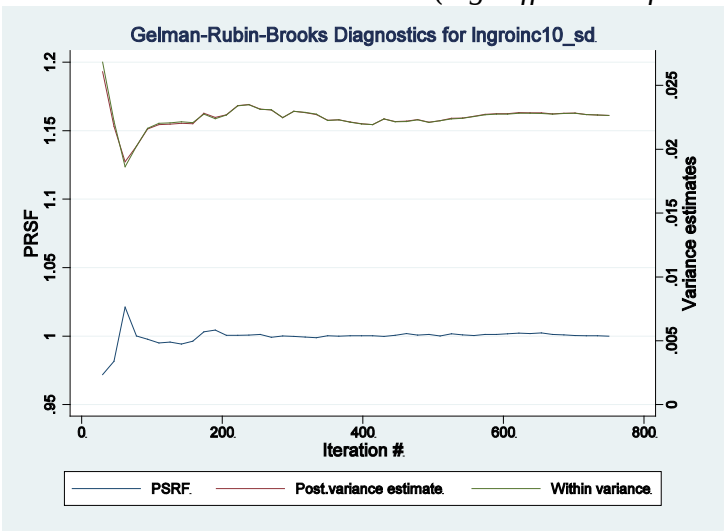


F10PGROINC

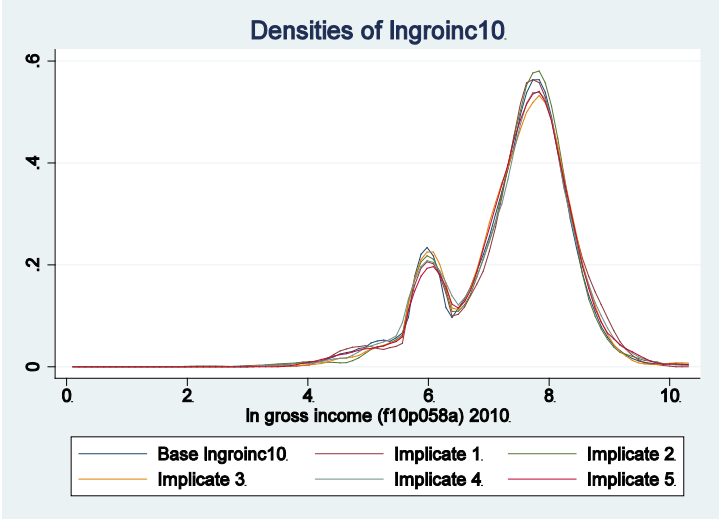
Gelman-Rubin Brooks Criteria: Mean (log-difference of net and gross income)



Gelman-Rubin Brooks Criteria: Standard Deviation (log-difference of net and gross income)

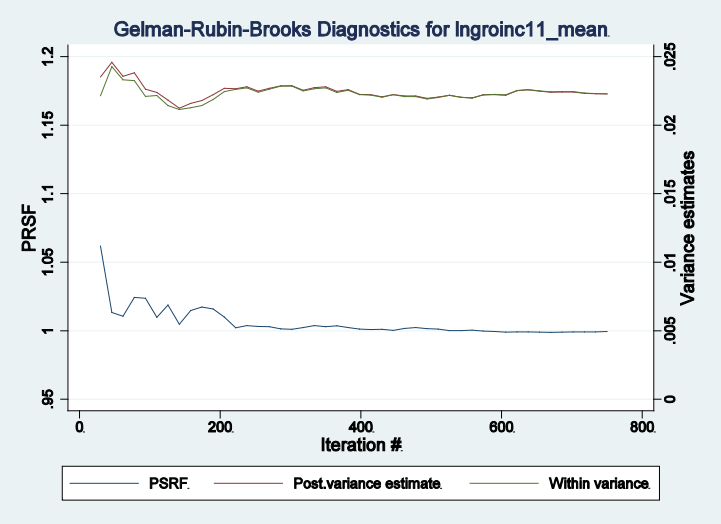


Comparison of Densities between original distribution (blue) and the five implicates

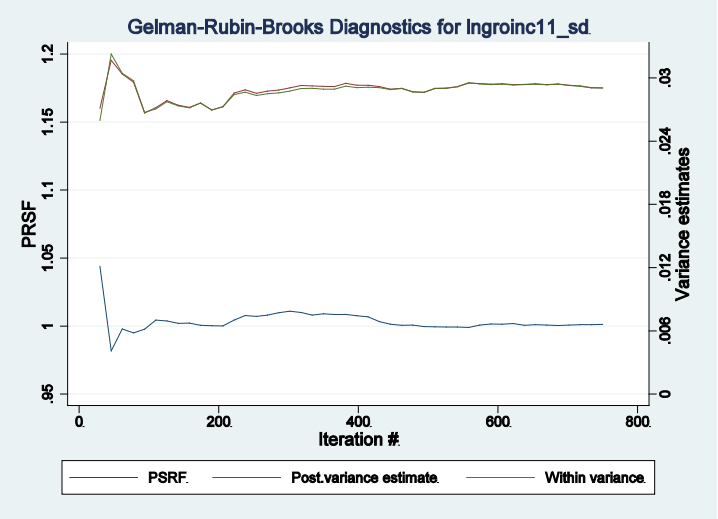


F11PGROINC

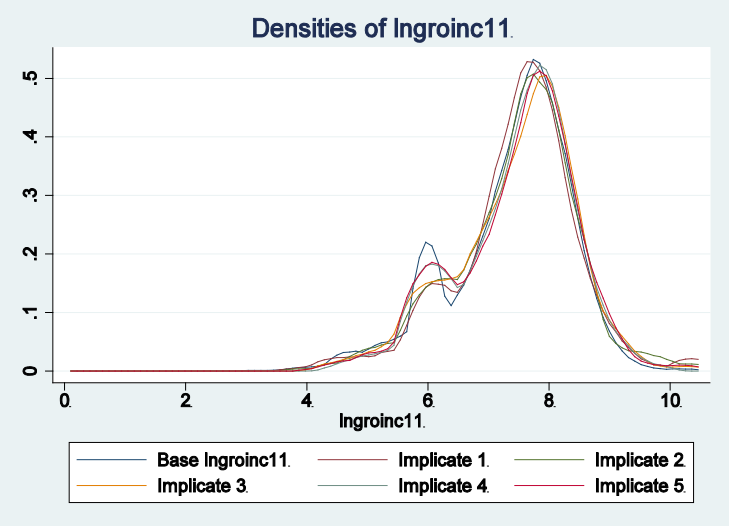
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

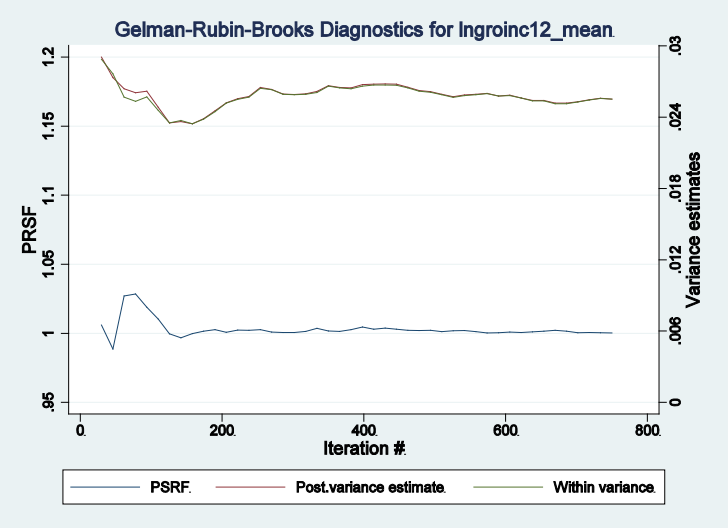


Comparison of Densities between original distribution (blue) and the five implicates

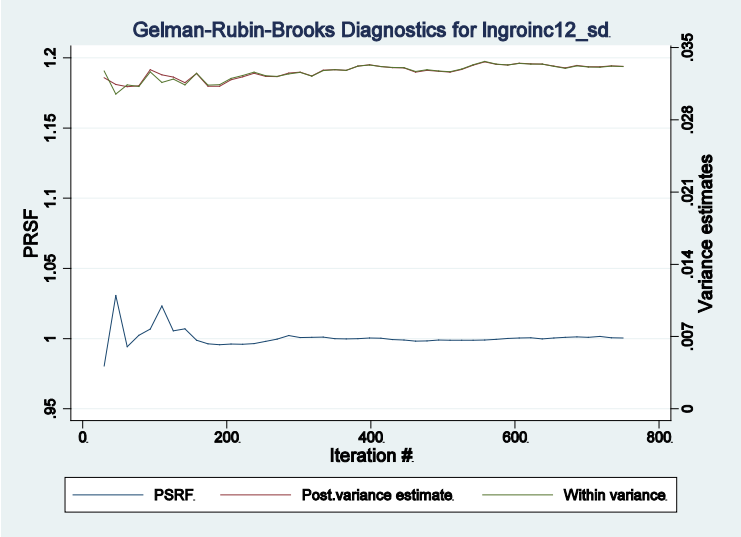


F12PGROINC

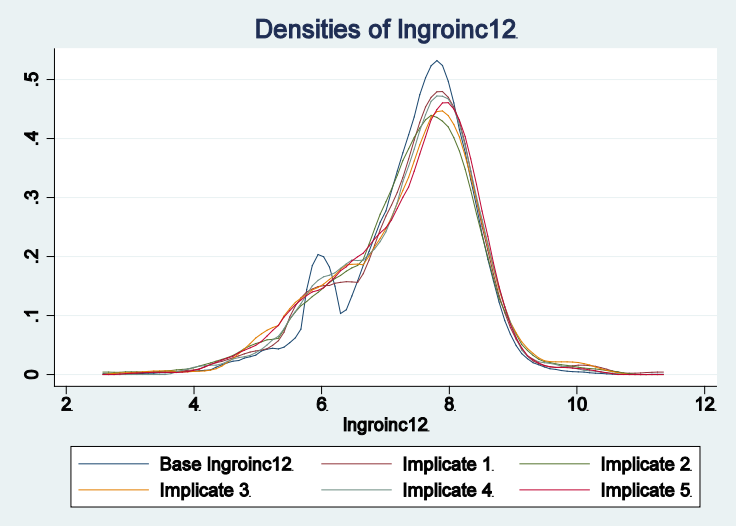
Gelman-Rubin Brooks Criteria: Mean (log-difference of net and gross income)



Gelman-Rubin Brooks Criteria: Standard Deviation (log-difference of net and gross income)

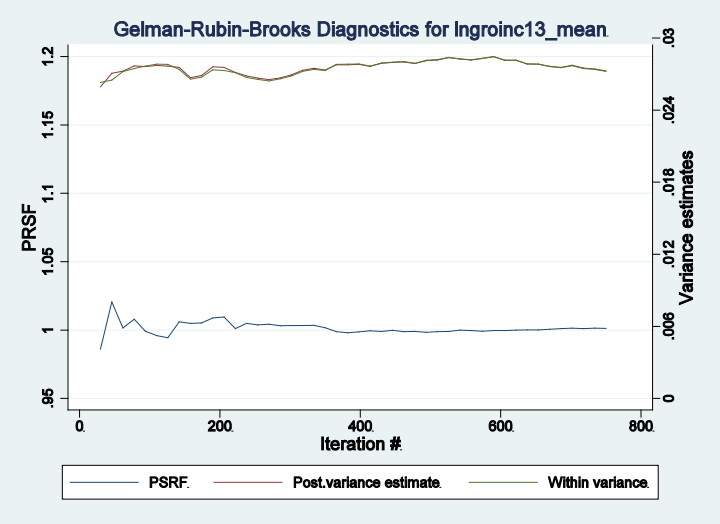


Comparison of Densities between original distribution (blue) and the five implicates

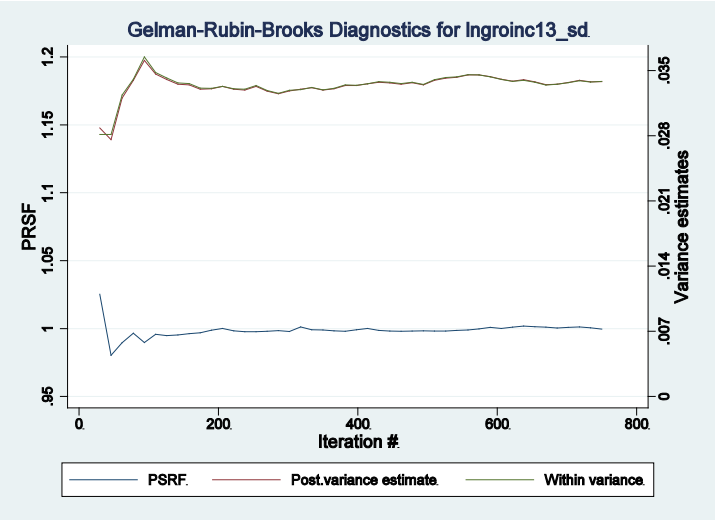


F13PGROINC

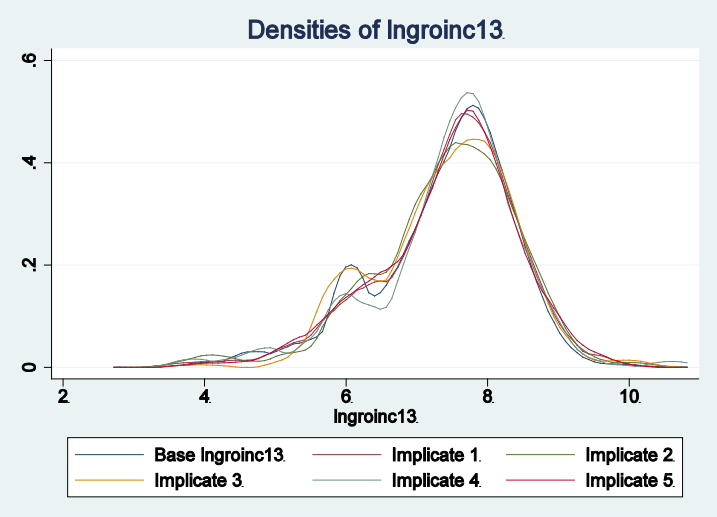
Gelman-Rubin Brooks Criteria: Mean (log-difference of net and gross income)



Gelman-Rubin Brooks Criteria: Standard Deviation (log-difference of net and gross income)



Comparison of Densities between original distribution (blue) and the five implicates



Variables in $mihinc

_MI

Variable label **"observation number"**
Variable format 4-digit integer

Comment _mi identifies observations belonging to the same household. For each household, there are 6 observations – one original and 5 imputed values.

_MJ

Variable label **"imputation number"**
Variable format 1-digit integer

Comment _mj identifies the implicates for each household and variable. The range is from 0 to 5, where 0 denotes the original observation and 1-5 identify the respective imputations.

$HHINC

Variable label **"monthly net household income"**
Variable format 5-digit integer

Comment This variable is the so-called income "screener" and records the overall household net income at the current interview month. Note that all households are "eligible" to have a household income, hence the code "-2" (does not apply) is only set for this variable in the original variable (where _MJ=0). The natural logarithm of this variable was imputed, and then converted back to the monetary values.

$HCHDBEN

Variable label **"monthly child benefits ('Kindergeld')"**
Variable format 4-digit integer

Comment This variable originates from the current report of child benefits. Note that "-2" (does not apply) is set for all households which do not receive child benefits ("Kindergeld").

$HCHDADD

Variable label **"monthly added child benefits ('Kinderzuschlag')"**
Variable format 3-digit integer

Comment The "Kinderzuschlag" question concerns the specific benefits given to households with low income, which are not receiving any ALG2 benefits. It is not quite clear, whether individuals have correctly understood this question, as the name of the benefit is somewhat ambiguous. Note that "-2" (does not apply) is set for all households

which do not receive the added child benefit for low income households.

$HUEBEN2

Variable label **"monthly long-term UE benefits ('ALG2')"**
Variable format 4-digit integer

Comment This variable is based on the household receipt of long term unemployment benefits ("ALG2" or "Hartz IV"). Note that "-2" (does not apply) is set for all households which do not receive any long-term unemployment benefits.

$HCARBEN

Variable label **"monthly care benefits 'PflegeVers' (imputed)"**
Variable format 3-digit integer

Comment This variable covers the imputations for the household care benefits ('Pflegeversicherung'). Note that "-2" (does not apply) is set for all households which do not report any care benefits.

$HHELBEN

Variable label **"monthly other help benefits 'Hilfe Lebenslagen' (imputed)"**
Variable format 3-digit integer

Comment This variable covers the imputations for the household's other help benefits ('Hilfe in besonderen Lebenslagen'). Note that "-2" (does not apply) is set for all households which do not report to receive any of these benefits.

$HAGETRN

Variable label **"monthly age transfer benefit 'Grundsicherung Alter' (imputed)"**
Variable format 3-digit integer

Comment This variable covers the imputations for the household receipt of age transfer benefits ('Grundsicherung im Alter und bei Erwerbsminderung'). Note that "-2" (does not apply) is set for all households which do not report to receive any of these benefits.

$HHOSBEN

Variable label **"monthly housing benefits ('Wohngeld')"**
Variable format 3-digit integer

Comment The origin is in the question about housing benefits received by the household ("Wohngeld"). Note that "-2" (does not apply) is set for all households which do not receive any housing benefits.

$HRENT

Variable label **"monthly rent payments"**
Variable format 4-digit integer

Comment The origin of this variable is the question about rent payments. As concerns the imputation, there is no distinction between gross and net rent, although the information, whether utilities are included in the rent payments is part of the imputation model and thus is accounted for. Note that "-2" (does not apply) is set for all owner-occupier households and renter-occupier households which do not pay rent. Note also that this variable *does not* correspond to the variable rent$$ in ***$hgen***.

$HUTIL

Variable label **"monthly utility payments"**
Variable format 3-digit integer

Comment The variable originates in the report of the monthly utility costs. There is no distinction between utility costs which are partially or fully included in the rent, although a corresponding indicator variable is included in the imputation process. Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent or any utilities.

$HHEAT

Variable label **"monthly heating payments"**
Variable format 3-digit integer

Comment The variable originates in the report (or the respondent's estimate) of the monthly heating costs. Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent.

$HELEC

Variable label **"monthly electricity payments"**
Variable format 3-digit integer

Comment Electricity payments are derived from the report of each renter on these costs. Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent.

$HCREDIT

Variable label **"monthly credit payments"**
Variable format 3-digit integer

Comment This variable covers the imputations for the household credit payments which are not mortgages. Note that "-2" (does not apply) is set for all households which do not report any credit burden.

I_$HHINC

Variable label **"$hhinc is imputed"**
Value labels I_$HHINC
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that all households are "eligible" to have a household income, hence the code "-2" (does not apply) is only set for this variable in the original variable (where _MJ=0).

I_$HCHDBEN

Variable label **"$hchdsup is imputed"**
Value labels I_$HCHDBEN
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all households which do not receive child benefits ("Kindergeld").

I_$HCHDADD

Variable label **"$hchdadd is imputed"**
Value labels I_$HCHDAAD
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all households which do not receive the added child benefit for low income households ("Kinderzuschlag").

I_$HUEBEN2

Variable label **"$hueben is imputed"**
Value labels I_$HUEBEN2
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all households which do not receive any long-term unemployment benefits ("ALG2").

I_$HCARBEN

Variable label	**"$hcarben is imputed"**
Value labels	I_$HCARBEN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all households which do not receive any care benefits ("Pflegeversicherung").

I_$HHELBEN

Variable label	**"$hhelben is imputed"**
Value labels	I_$HHELBEN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all households which do not receive any help benefits ("Hilfe in besonderen Lebenslagen").

I_$HAGETRN

Variable label	**"$hagetrn is imputed"**
Value labels	I_$HAGETRN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all households which do not receive age transfer benefits ('Grundsicherung im Alter und bei Erwerbsminderung')

I_$HHOSBEN

Variable label	**"$hhosben is imputed"**
Value labels	I_$HHOSBEN (0) Observed value (1) Imputed value
Variable format	1-digit integer
Comment	Note that "-2" (does not apply) is set for all households which do not receive any housing benefits ("Wohngeld").

I_$HRENT

Variable label	**"$hrent is imputed"**
Value labels	I_$HRENT (0) Observed value (1) Imputed value

Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all owner-occupier households and renter-occupier households which do not pay rent.

I_$HUTIL

Variable label **"$hutil is imputed"**
Value labels I_$HUTIL
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent or any utilities.

I_$HHEAT

Variable label **"$hheat is imputed"**
Value labels I_$HHEAT
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent.

I_$HELEC

Variable label **"$helec is imputed"**
Value labels I_$HELEC
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all owner-occupier households and households reporting not to pay rent.

I_$HCREDIT

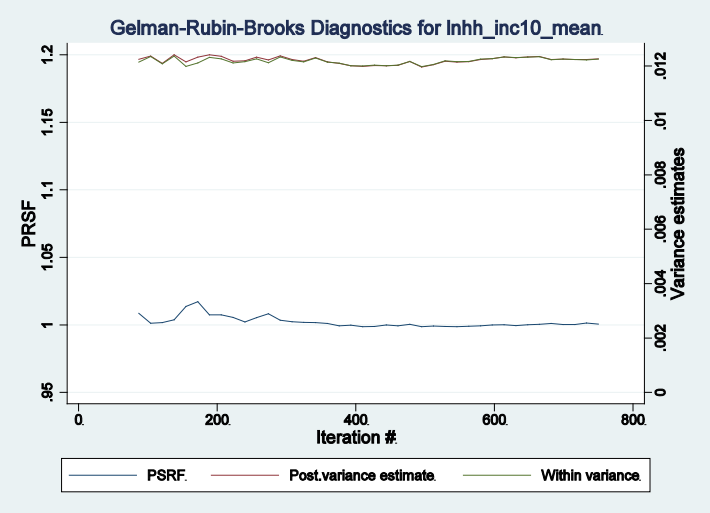
Variable label **"$hcredit is imputed"**
Value labels I_$HCREDIT
(0) Observed value
(1) Imputed value
Variable format 1-digit integer

Comment Note that "-2" (does not apply) is set for all households which do not report any credit burden.

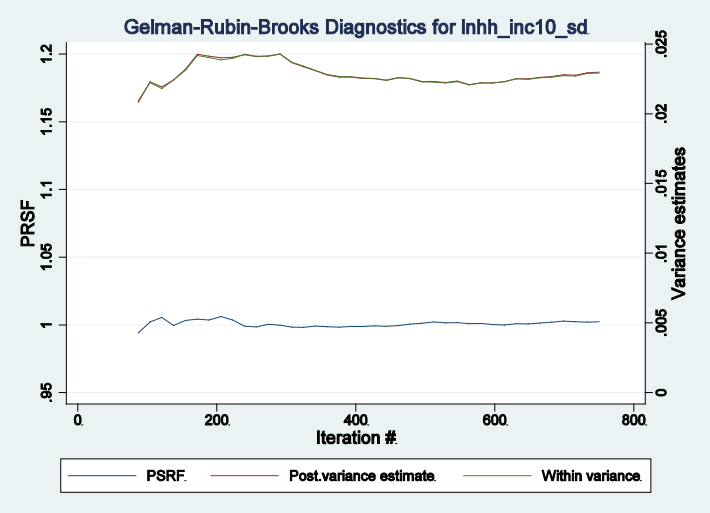
Evaluation graphs for $mihinc

F10HHINC

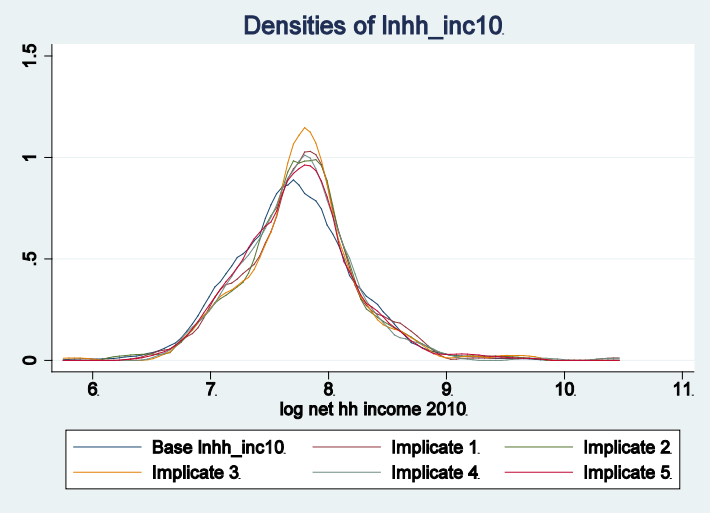
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

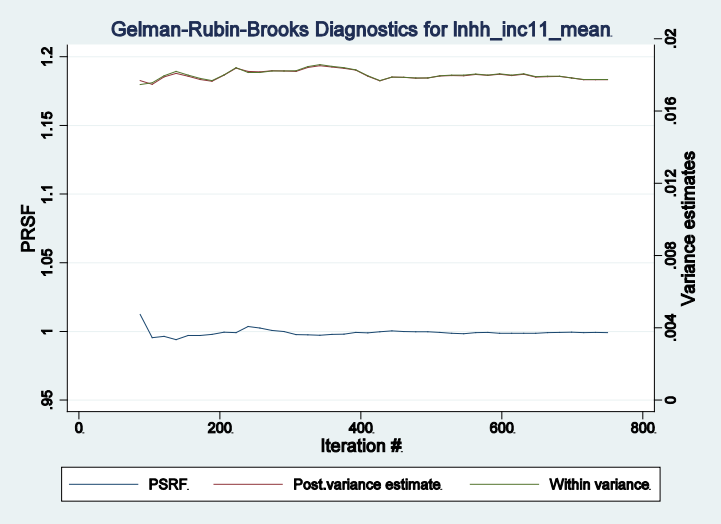


Comparison of Densities between original distribution (blue) and the five implicates

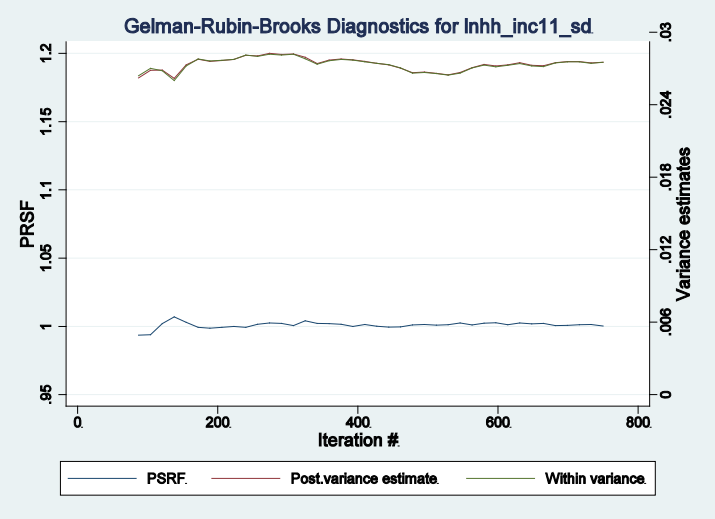


F11HHINC

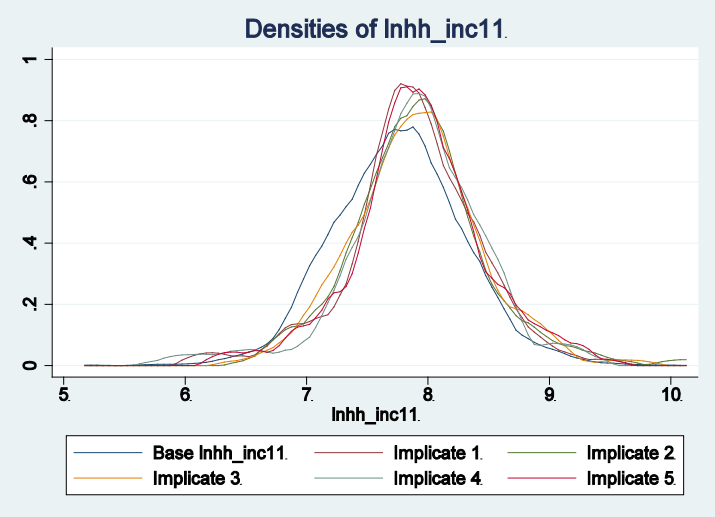
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

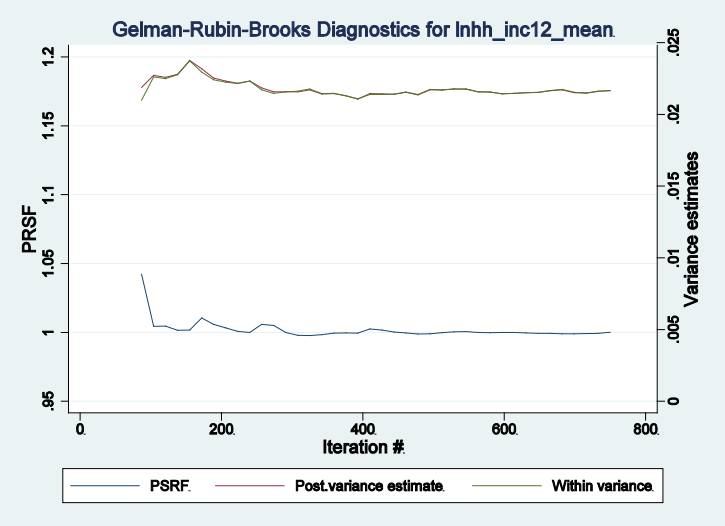


Comparison of Densities between original distribution (blue) and the five implicates

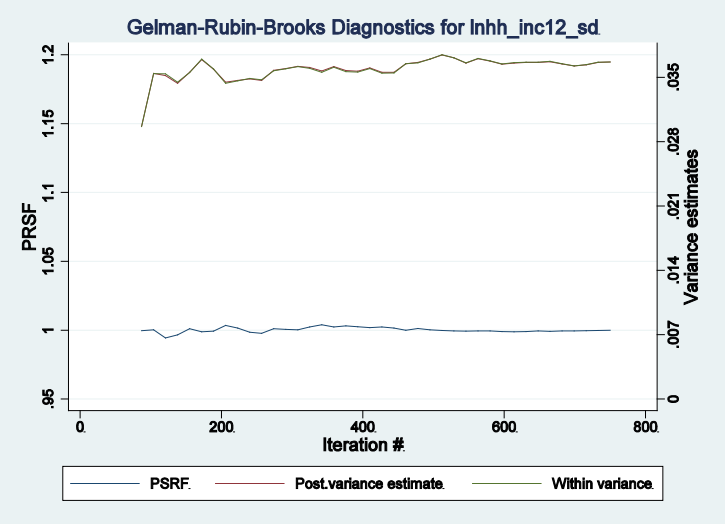


F12HHINC

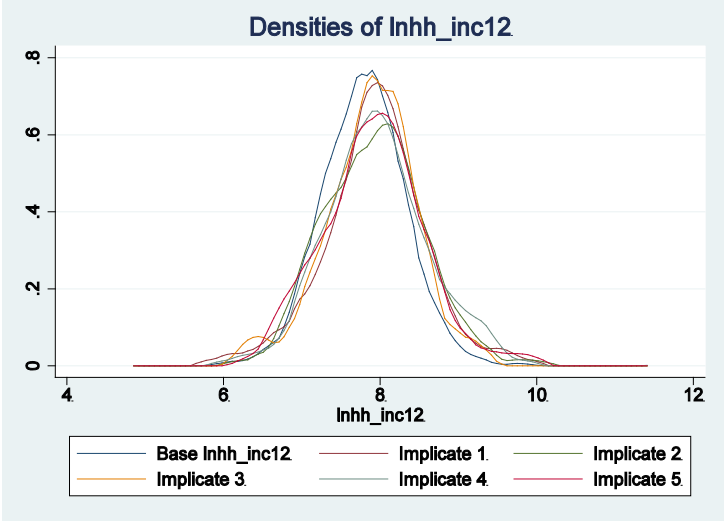
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation

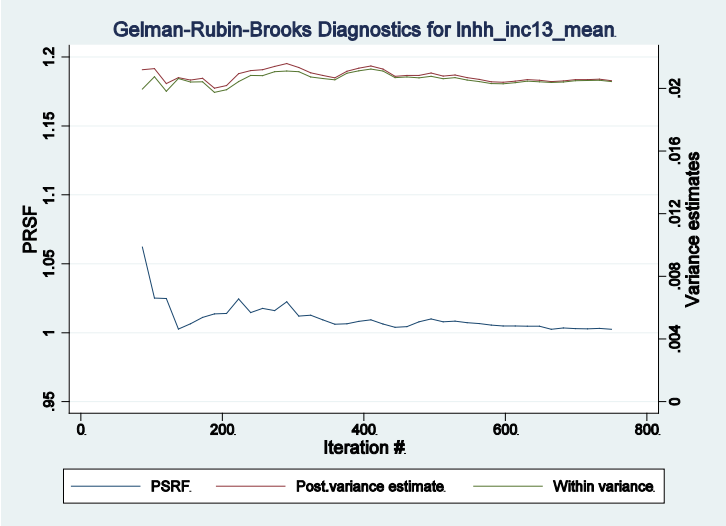


Comparison of Densities between original distribution (blue) and the five implicates

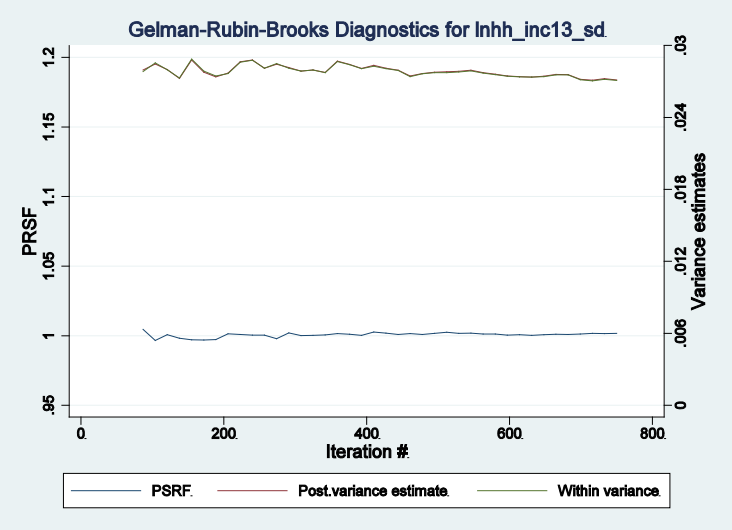


F13HHINC

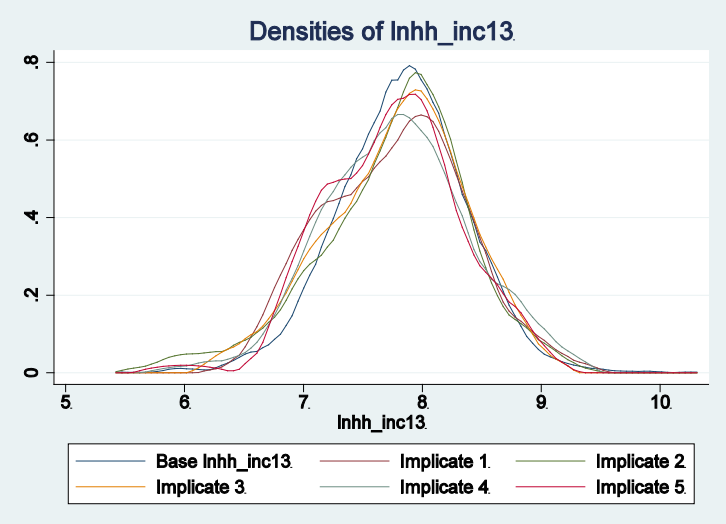
Gelman-Rubin Brooks Criteria: Mean



Gelman-Rubin Brooks Criteria: Standard Deviation



Comparison of Densities between original distribution (blue) and the five implicates



References

Brooks, S.P. and A. Gelman (1998): General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7 (4), pp 434-455.

Little, R.J.A. (1992): Regression with Missing X's: A Review. *Journal of the American Statistical Association*, 87 (420), pp 1227-1237.

Royston, P. (2004): Multiple imputation of missing values. Stata Journal 4(3): 227-241.

Royston, P. (2005a): Multiple imputation of missing values: update. *Stata Journal* 5(2): 188-201.

Royston, P. (2005b): Multiple imputation of missing values: Update of ice. *Stata Journal* 5(4): 527-536.

Royston, P. (2007): Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 7(4): 445-464.

Royston, P. (2010): *ice.ado*. Version 1.9.1 PR/IW 20aug2010.

Rubin, D.B. (1987): Multiple imputation for non-response in surveys. New York.

Rubin, D.B. (1996): Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91 (434), pp 473-488.

Schafer, J.L. and J.W.Graham (2002): Missing Data: Our view of the state of the art. *Psychological Methods*, 7(2), pp 147-177.

Starick, R. and N. Watson (2007): Evaluation of alternative income imputation methods for the HILDA Survey. *HILDA Project Discussion Paper Series*, No 1/07.

StataCorp. (2013). Stata multiple imputation reference manual, Release 13. Statistical Software. College Station, TX: StataCorp LP.

Van Buuren, S, J.P.L.Brands, C.G.M.Groothuis-Oudshoorn and D.B.Rubin (2006): Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76 (12), pp 1049-1064.

Documentation *pbiospe*

Activity Biography

Stefan Damerow

*This documentation is based on the comparable SOEP documentation on **pbiospe** and has benefited from previous work by Henning Lohmann. Please understand that for readability reasons, we do not specifically cite and specify text that has been used directly from the SOEP document.*

pbiospe

The spell file ***pbiospe*** is based on the information on activity status over the life course, which is collected as a matrix from every respondent using the Biography Questionnaire (in FiD, this belongs to the second part of the biography, i.e. ***$lela*** with $LELTYP=2 or 3). The observations start at the age of 15 and end at the current age (up to age 65). The information on the activity status covers only the period up to the time the biography is collected. To update the ongoing occupational career in ***pbiospe***, information from the yearly Person Questionnaire is also used. In this questionnaire, respondents are always asked their occupational status for every month of the previous year. Therefore, the information collected on a monthly basis and stored in the file ***artkalen*** is aggregated into yearly values and combined with the information gathered from the Biography Questionnaire.

In the following, the method of combining the data is described. First of all we provide a brief overview of the contents of ***pbiospe***. Table 1 contains a list of all the variables in the dataset. The variables BEGIN and END indicate the beginning and the end of a spell. These variables are age entries. There are also variables that refer to calendar years: BEGINY and ENDY (Y stands for Year). The variable SPELLTYP contains information on the activity status during the spell, e.g., employed full-time or unemployed. The SPELLNR is a serial identifier of spells of a given person. Missing information on the beginning or end of a spell cause what are known as censoring problems. There are two types of missing data. First, data can be missing on periods outside the observation window (before the age of 15 and after the age of 65). Second, data can be missing on years within the observation window due to item non-response in particular years or due to temporary drop-outs (the latter applies to calendar information only). In this case, we speak of “gaps.” There are nine different patterns (variable ZENSOR):

1. uncensored: beginning observed, end observed
2. right-censored: beginning observed, end not observed
3. right-censored (gap): beginning observed, end not observed because of gap
4. left-censored: beginning not observed, end observed
5. left- and right-censored: beginning not observed, end not observed
6. left-censored and right-censored (gap): beginning not observed, end not observed because of gap
7. left-censored (gap): beginning not observed because of gap, end observed

8. left-censored (gap) and right-censored: beginning not observed because of gap, end not observed
9. left-censored (gap) and right-censored (gap): beginning not observed because of gap, end not observed because of gap

Table 1: Contents of *pbiospe* (variables)[29]

hhnr	Original Household Number
persnr	Never Changing Person ID
spellnr	Serial Number of the Event per Person
spelltyp	Type of Event
begin	Age spell begins
end	Age spell ends
beginy	Year spell begins
endy	Year spell ends
zensor	Zensor Variable
spellinf	Spell construction information
erhebj	Survey year biography data
kalyear	First observation year calendar
beginb1	Age spell begins, 1. initial biography spell
endb1	Age spell ends, 1. initial biography spell
begink1	Age spell begins, 1. initial calendar spell
endk1	Age spell ends, 1. initial calendar spell
beginyb1	Year spell begins, 1. initial biography spell
endyb1	Year spell ends, 1. initial biography spell
beginyk1	Year spell begins, 1. initial calendar spell
endyk1	Year spell ends, 1. initial calendar spell
beginb2	Age spell begins, 2. initial biography spell
endb2	Age spell ends, 2. initial biography spell
beginyb2	Year spell begins, 2. initial biography spell
endyb2	Year spell ends, 2. initial biography spell
begink2	Age spell begins, 2. initial calendar spell
endk2	Age spell ends, 2. initial calendar spell
beginyk2	Year spell begins, 2. initial calendar spell
endyk2	Year spell ends, 2. initial calendar spell

As mentioned above, ***pbiospe*** combines information collected in the biography questionnaire and the calendar matrix of the individual questionnaire. The two types of information are merged into ***pbiospe*** following a number of rules. First of all, it is important to acknowledge that the Biography Questionnaire Matrix as well as the Individual Questionnaire Matrix allow

[29] In SOEP the data file ***pbiospe*** also contains the variables beginb3 – endyk4. These additional variables become relevant for longer panel durations only (for further explanation see below).

for multiple activity statuses for a given year or month. No concept of main activity is used. A common combination is, for instance, “housewife/-husband” and “working part-time”. There are a number of other plausible combinations, but also combinations that are less plausible. However, a list of valid combinations of activity statuses defined according to legal or similar constructs would need to be based on very strong assumptions. In addition—in particular in case of the yearly matrix in the Biography Questionnaire—activities are reported that took place in a calendar year in consecutive months, which makes it impossible to exclude combinations of activities. Therefore, no data cleaning is performed at this stage. As a consequence, the data may contain information on more than one activity for a given point in time.

This also defines the rules for aggregating the monthly ***artkalen*** data into yearly values. Take, for example, a person who was in full-time employment from January to November 2007, and unemployed in December 2007. The exact months are recorded in the dataset ***artkalen***. In the aggregated data, which is merged with the yearly data from the Biography Questionnaire, you find the information that the person worked full-time and was also unemployed in the year 2007. There is a second level of aggregation of ***artkalen*** information as the data on type of activity, which is recorded in the variable SPELLTYP is more detailed than in ***pbiospe***. The respective information is aggregated as described in Table 2.

Table 2: Aggregation of *artkalen* spell information into *pbiospe*

	pbiospe	***artkalen***
1	School/University	School, College (1)
2	Apprenticeship/Training	Vocational Training (4), First Job Training, Apprenticeship (13), Continuing Education, Retraining (14)
3	Military/Civilian service	Military, Community Service (9)
4	Full-time employed	Full-Time Employment (1), Short Work Hrs (2)
5	Part-time employed	Part-Time Employment (3), Second Job (11), Mini-job (up to 400 euros) (15)
6	Unemployed	Unemployed (5)
7	House-Husband/Wife	Housewife, Husband (10)
8	Retired	Retired (6)
9	Other	Maternity Leave (7), Other (12)
99	Gap	Information on gaps in ***artkalen*** is not used. Gaps are calculated on the basis of the merged dataset.

After merging the information from the Biography Questionnaire and ***artkalen***, the data are transformed into spells, whereby each spell is defined by the duration of a given status. A

question that arises when merging the data is how to handle overlapping pieces of information. The basic principle is to assign a value of a given status in a given year if the status is recorded in the calendar or in the biography information or both. An example might help to illustrate this: the calendar records full-time employment for the years 2010 and 2011 while the biography records full-time employment for the period from 2000 up to 2011. The merged data from ***pbiospe*** contains a spell that begins in 2000 and ends in 2011. However, the initial information is restored by including additional variables, which allows for alternative ways of merging the data (see below). The variables SPELLINF, ERHEBJ, and KALYEAR contain general information on the sources of the information captured in a given spell. Table 3 shows that the majority of spells are based on biography information only (64.3 percent). Almost one sixth of all spells (15.5 percent) are not observed in the Biography Questionnaire but only in the calendar data. The remainder of spells contains information from biography as well as calendar data. Usually these spells combine one period observed in the Biography Questionnaire with a period observed in the calendar.

Table 3: Sources of *pbiospe* spells

Spell construction information	n	%	cum. %
[-2] Does not apply	118	0.22	0.22
[1] Biography only	32,606	61.75	61.97
[2] Calendar only	9,809	18.58	80.54
[3] 1 Biography + 1 Calendar spell	9,950	18.84	99.39
[4] 2+ Biography and 1 Calendar spell(s)	164	0.31	99.7
[5] 1 Biography and 2+ Calendar spell(s)	150	0.28	99.98
[6] 2+ Biography and 2+ Calendar spell(s)	10	0.02	100
Total	52,807	100	

Source: FiDv4.0

The variables BEGINB1-ENDYB2[30] document the initial information from the two different sources and are probably not of interest to the majority of users. However, on the basis of these variables, users are able to fully separate the Biography data from the aggregated ***artkalen*** data. This is advisable if you want to use the more detailed ***artkalen*** information and combine it with the yearly information from ***pbiospe*** for earlier years only. The variable names indicate the "source" of the original information utilized (B: Biography -Questionnaire

[30] Again, the variety of combinations of biography and calendar spells will increase in upcoming waves.

or K: calendar information from the yearly survey). As an example from SOEP v27, we discuss one of the spells that combines information on more than one period from any of the two sources. The spell number 4 of person 9205 starts in 1983 and ends in 1994 (SPELLTYP=4: full-time employment). As the variable SPELLINF (=5) shows, this a spell that combines one period from the biography data with two periods from the calendar data. According to the biography data, the person worked full-time from 1983 (BEGINYB1) until 1992 (ENDYB1). There is overlapping information from the calendar data available from 1986 onwards (KALYEAR). According to these data, the person worked full-time from 1986 (BEGINYK1) to 1990 (ENDYK1) and from 1993 (BEGINYK2) to 1994 (ENDYK2). During the years 1991 and 1992, no full-time employment is recorded in the calendar data, which contradicts the information from the biography data.

Table 5: Example of combined spell

persnr	spellnr	spelltyp	beginy	endy	spellinf	erhebj	kalyear	beginyb1	endyb1	beginyk1	endyk1	beginyk2	endyk2
9205	4	4	1983	1994	5	1998	1986	1983	1992	1986	1990	1993	1994

Source: SOEP v27 (***pbiospe***).

In ***pbiospe***, no attempt is made to "resolve" such contradictions, as this would require rather strong assumptions. More important, such assumptions would differ according to the research question, which makes it even more difficult to provide a standard solution. Therefore, in such cases, we generate spells in the same manner as in less difficult cases, namely by combining the information from the calendar and the biography data. However, users who are interested in combining biography and calendar data in a different manner can use the variables BEGINB1-ENDYB2 to fully separate the two types of data and to recombine the data on the basis of different rules of aggregation.

Documentation *pbrutto*

Gross information on all persons in the household

Note that part of this information is in German, based on the codebook provided by TNS-Infratest for FiD and SOEP.

List of variables:

$GEBURT

Variable Label	**„Geburtsjahr“**
Value Labels	$GEBURT
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13

Comment: Contains the year of birth the sample member. Is used to determine how the sample member should be treated (e.g. new or old respondent). See also $BEFSTAT.

$SEX

Variable Label	**„Geschlecht“**
Value Labels	$SEX (1) männlich (2) weiblich
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13

Comment: Respondent gender. The code “9 unbekannt” is not set in FiD.

$PNAT

Variable Label: **„Staatsangehörigkeit“**
Value Labels: $PNAT

(001) Deutschland
(002) Türkei
(003) Jugoslawien/ Montenegro/ Serbien
(004) Griechenland
(005) Italien
(006) Spanien
(007) Ex-DDR
(010) Österreich
(011) Frankreich
(012) Benelux
(013) Dänemark
(014) Großbritannien
(015) Schweden
(016) Norwegen
(017) Finnland
(018) USA
(019) Schweiz
(020) Chile
(021) Rumänien
(022) Polen
(023) Korea
(024) Iran
(025) Indonesien
(026) Ungarn
(027) Bolivien
(028) Portugal
(029) Bulgarien
(030) Syrien
(031) Tschechien
(032) Russland
(034) Mexiko
(035) Argentinien
(036) Cap Verde
(037) Benin
(038) Philippinen
(039) Israel
(040) Japan
(041) Australien
(042) Indien
(043) Afghanistan
(044) Thailand
(045) Jamaika
(046) Saudi Arabien
(047) Äthiopien
(049) Ghana
(050) Bangladesch
(051) Venezuela
(052) Tunesien
(053) Mauritius
(054) Nigeria
(055) Kanada
(056) Neuseeland
(057) Tansania
(058) Zypern
(059) Kuba
(060) Irak
(061) Brasilien
(062) Monaco
(063) Hongkong
(064) Peru
(065) SriLanka
(066) Nepal
(067) Marokko
(068) China
(069) Liechtenstein
(070) Island
(071) Irland
(072) St.Lucia
(073) Moldawien
(074) Kasachstan
(075) Albanien
(076) Libanon
(077) Kirgistan
(078) Ukraine
(079) Algerien
(080) Mocambique
(081) Ägypten
(082) Tadschikistan
(083) Vietnam
(084) Somalia
(085) Pakistan
(086) Südafrika
(087) Ver.Arabische Emirate
(088) ElSalvador
(089) Eritrea
(090) Jordanien
(092) CostaRica
(093) Singapur
(094) Burkina Faso
(095) Sambia
(096) Ecuador
(097) Usbekistan
(098) Staatenlos
(099) Puerto Rico
(100) Laos
(101) Estland
(102) Angola

(103) Lettland
(105) Namibia
(107) Belize
(108) Dominikanische Republik
(109) Nicaragua
(110) Kenia
(111) Libyen
(112) Malta
(113) Botswana
(114) Haiti
(115) Trinidad, Tobago
(116) Luxemburg
(117) Belgien
(118) Holland
(119) Kroatien
(120) Bosnien-Herzegowina
(121) Makedonien
(122) Slowenien
(123) Slowakei
(124) Paraquay
(125) Guinea
(126) Kuwait
(127) Elfenbeinküste
(128) Malaysia
(129) Samoa
(130) Aserbaidschan
(131) Seychellen
(132) Weißrußland
(133) Uruguay
(134) Bahamas
(135) Uganda
(136) Oman
(137)Mikronesien
(138) Mali
(139) Kamerun
(140) Kosovo-Albaner
(141) Georgien
(142) Sudan
(143) Kongo
(144) Togo
(145) Mongolei
(146) Litauen
(147) Tschad
(148) Armenien
(149) Kurdistan
(150) Liberia
(151) Jemen
(152) Palästina
(153) Freistaat Danzig
(154) Taiwan
(155) Turkmenistan
(156) Afrika
(157) Guatemala
(158) Sierra Leone (Westafrika)
(159) Panama
(160) Osttimor
(161) Bahrain
(162) Senegal
(163) Malediven
(164) Hawaii
(165) Serbien
(166) Gambia
(167) Honduras
(168) Montenegro
(169) Kambodscha
(170) Surinam
(171) Guyana
(172) Kaukasus
(173) Simbabwe
(174) Madagaskar
(175) Grenada

Variable format 3-digit integer
$ - Wave F10, F11, F12, F13

Comment Respondent's nationality.

$STISTAT

Variable Label **„Stichprobenstatus“**
Value Labels $STISTAT
(1) Stichprobenmitglied Stammperson alt
(2) Nicht-Stichprobenmitglied (zugeordnete Person) alt
(3) Stichprobenmitglied neu
(4) Nicht-Stichprobenmitglied neu
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Sample Status
A person's sample status "Stichprobenstatus“ is a fixed person characteristic given at the first contact with the study. All individuals living in the household in the first wave are sample persons. Individuals entering the study in any following wave are usually non-sample members. Children born into an old panel household usually receive code 3 and are thus considered panel members. If a family member returns from abroad, she is also considered a panel member and receives cod 3. All other persons receive code 4 (usually individuals moving into a household from within Germany and/or children from non-sample members. In a person's initial wave, code 3 or 4 is given, which in the following wave is changed to code 1 or 2.

$BEFSTAT

Variable Label **„Befragungsstatus“**

Value Labels	$BEFSTAT (1) Erneut zu befragen (2) Erstmals zu befragen, da bisher temporärer Ausfall (3) Erstmals zu befragen, da Befragungsalter erreicht (4) Erstmals zu befragende neue Person (5) Noch nicht zu befragende alte Person (6) Noch nicht zu befragende neue Person (7) Nicht zu befragen, da harter Verweigerer in Vorwelle/n (8) Umwandlung von Code 7, da doch wieder teilgenommen
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Respondent's question status „Befragungsstatus" allows the survey agency (and thus the interviewer) to know in advance, what type of interview (if any) is to be expected in this wave. Codes 1, 2, 3, 5, 7 are set before the field period starts based on information from the previous years. These codes are not changed during the field work, even if a household member moves into a new household. [31] Codes 4 and 6 are set manually for all new individuals (also for those in new samples) considering the following rules: Code 4 is set, when the person is eligible in terms of age, i.e. 17 years old or older. Code 6 is set whenever individuals are not age eligible. Code 8 is given only for the case when a household member refused explicitly to respond in a previous wave. These individuals receive a code 7 initially. However, in some cases respondents change their mind in a following wave and decide to participate. In that case, they receive the code 8, removing the prior code 7.

$LINT

Variable Label	**„Letztes Interview in Welle"**
Value Labels	$LINT (0) Befragungsperson, jedoch noch nie ein Personeninterview gegeben (2010) Letztes Interview in Welle 2010. (2011) Letztes Interview in Welle 2011. (2012) Letztes Interview in Welle 2012. (2013) Letztes Interview in Welle 2013.
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	This code is set after the end of the field work on the basis of the resulst of the interview. If an interview was conducted, the current year will appear in this wave.

[31] Refusing individuals of the previous year ($BEFSTAT=7) are not followed in case they move (see also $PADER=7). However, in case they move with a person still in the panel, a person dataset is created for them ($PADER=8).

$LUECKE

Variable Label **„Lückenbearbeitung“**
Value Labels $LUECKE
(6) Lücke ist entstanden, da Person in aktueller Welle wieder teilgenommen hat.
(7) Keine Lücke entstanden, da Person auch in aktueller Welle nicht teilgenommen hat.
(8) Lücke über mehrere Vorwellen
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable documents whether a gap exists in the person’s panel life. Note that for most persons, this variable takes the code “-2 does not apply”, as they participated in a previous wave.

$ZUPAN

Variable Label **„Zugangswelle zum Panel“**
Variable format 4-digit integer
$ - Wave F10, F11, F12, F13

Comment $ZUPAN shows in which wave a person entered the panel, i.e. is first listed in the gross sample. This variable is thus constant over time.

$PNRAKT

Variable Label **„Aktuelle Personennummer“**
Variable format 2- digit integer
$ - Wave F10, F11, F12, F13

Comment This variable depicts the order in which the household members are listed in the address protocol, where the order is set usually when the interviewer enters the household. This number is a useful identifier in cases the PERSNR of a person answering the questionnaire is unknown in case of a proxy interview (i.e. in the household files, ***$h***, $PNRAKT allows to identify the person answering the questionnaire via the corresponding variable $AUSKU).

$STELL

Variable Label **„Stellung zum Haushaltsvorstand“**
Version f10, f11
Value Labels $STELL

(00)	Haushaltsvorstand
(01)	Ehepartner des HV
(02)	Lebenspartner des HV
(03)	Kind (auch Adoptivkind) HV
(04)	Pflegekind des HV
(05)	Schwiegersohn/-tochter HV
(06)	Eltern HV

(07)	Schwiegereltern HV
(08)	Geschwister und Schwager/Schwägerin HV
(09)	Enkel des HV
(10)	Onkel, Tante, Neffe, Nichte u.ä. HV
(11)	Nicht verwandte/verschwägerte Personen HV
(12)	Kind vom Lebenspartner des HV
(13)	gleichgeschlechtliche Ehepartner des HV nach dem Lebenspartnergesetz
(99)	Stellung zum HV unbekannt

$STELL (continued)

Version f12 and beyond

Value Labels $STELL

(00)	Haushaltsvorstand
(11)	Ehegatte des HV
(12)	Gleichgeschl. Ehepartner
(13)	Lebenspartner des HV
(21)	Sohn,Tochter des HV
(22)	Stiefkind des HV
(23)	Adoptivkind des HV
(24)	Pflegekind des HV
(25)	Enkel des HV
(26)	Urenkel des HV
(27)	Schwsohn,-tocher HV
(31)	Vater, Mutter des HV
(32)	Stiefmutter -vater/Partner d. leibl. Elternteils HV
(33)	Adoptivmutter, -vater HV
(34)	Pflegemutter, -vater HV
(35)	Schwvater,-mutter HV
(36)	Grossmutter, -vater HV
(41)	Schwester, Bruder HV
(42)	Halbschwester, -bruder HV
(43)	Stiefschwester, -bruder HV
(44)	Adoptivschwester, -bruder HV
(45)	Pflegeschwester, -bruder HV
(51)	Schwaegerin/Schwager 1: Ehegatten oder Lebenspartner von Geschwistern HV
(52)	Schwaegerin/Schwager 2: Geschwister von Ehegatten oder Lebenspartner HV
(61)	Tante/Onkel des HV
(62)	Nichte/Neffe des HV
(63)	Cousine/Cousin des HV
(64)	Sonst.mit HV verw.
(71)	Mit HV nicht verw.
(99)	Stellung zu HV unbekannt

Variable format 2- digit integer

$ - Wave F10, F11, F12, F13

Comment All households have household head (Haushaltsvorstand), usually the person most equipped to answer questions about the financial situation

of the household. Whenever possible, the household head remains constant over the years, unless a move or death requires to define a new household head. In that case, relationships via $STELL need to be newly identified.
Since 2012, the $STELL variable is asked in more detail to allow for more precise definitions. Given the table below, the new version can be "translated" into the old one.

$STELL (continued)

$STELL in waves F10 and F11		**$STELL in wave F12 and F13**	
0	Haushaltsvorstand (HV)	0	Haushaltsvorstand
1	Ehepartner	11	Ehegatte des HV
2	Lebenspartner	13	Lebenspartner des HV
3	Kind (auch Adoptivkind) HV	21	Sohn,Tochter des HV
		23	Adoptivkind des HV
4	Pflegekind des HV	24	Pflegekind des HV
5	Schwiegersohn/-tochter HV	27	Schwsohn,-tocher HV
6	Eltern HV	31	Vater, Mutter des HV
		33	Adoptivmutter, -vater HV
		34	Pflegemutter, -vater HV
7	Schwiegereltern HV	35	Schwvater,-mutter HV
8	Geschwister und Schwager/Schwägerin HV	41	Schwester, Bruder HV
		42	Halbschwester, -bruder HV
		43	Stiefschwester, -bruder HV
		44	Adoptivschwester, -bruder HV
		45	Pflegeschwester, -bruder HV
		51	Schwaegerin/Schwager 1: Ehegatten oder Lebenspartner von Geschwistern HV
		52	Schwaegerin/Schwager 2: Geschwister von Ehegatten oder Lebenspartner HV
9	Enkel des HV	25	Enkel des HV
10	Onkel, Tante, Neffe, Nichte u.ä. HV	26	Urenkel des HV
		32	Stiefmutter -vater/Partner d. leibl. Elternteils HV
		36	Grossmutter, -vater HV
		61	Tante/Onkel des HV
		62	Nichte/Neffe des HV
		63	Cousine/Cousin des HV
		64	Sonst.mit HV verw.
11	Nicht verwandte/verschwägerte Personen HV	71	Mit HV nicht verw.
12	Stiefkind des HV	22	Stiefkind des HV
13	Gleichgeschl. Ehepartner	12	Gleichgeschl. Ehepartner
99	Unbekannt	99	Stellung zu HV unbekannt

$PZUG

Variable Label **„Zugehörigkeit der Personen zum Haushalt“**
Value Labels $PZUG

(00) Lebt noch im Haushalt
(01) Vorübergehend abwesend: Bundeswehr
(02) Vorübergehend abwesend: Ausbildung/Studium
(03) Vorübergehend abwesend: Beruf/Montage
(04) Vorübergehend abwesend: Krankenhaus/Kur
(05) Vorübergehend abwesend: Längere Zeit verreist
(06) Vorübergehend abwesend: Sonstiges
(07) Verstorben
(08) Verzogen, jedoch keine Weiterverfolgung
(09) Personen, die in Vorwellen irrtümlich als Haushaltsmitglieder geführt worden sind

Neue Personen im alten Haushalt

(11) Geboren
(12) Zuzug aus Westdeutschland
(13) Zuzug aus dem Ausland
(14) In den Haushalt zurückgekehrt
(15) Zuzug aus dem Inland vor letzter Befragung und erstmals genannt
(16) Zuzug aus dem Ausland vor letzter Befragung und erstmals genannt
(17) Geboren vor letzter Befragung/schon immer im Haushalt gelebt und erstmals genannt
(18) Zuzug aus Ostdeutschland
(19) Keine Information

Alte Personen im neuen Haushalt

(20) Lebt im neuen Haushalt
(21) Vorübergehend abwesend: Bundeswehr
(22) Vorübergehend abwesend: Ausbildung/Studium
(23) Vorübergehend abwesend: Beruf/Montage
(24) Vorübergehend abwesend: Krankenhaus/Kur
(25) Vorübergehend abwesend: Längere Zeit verreist
(26) Vorübergehend abwesend: Sonstiges
(27) Verstorben
(28) Verzogen, jedoch keine Weiterverfolgung

Neue Personen im neuen Haushalt

(30) Lebt in diesem Haushalt schon länger
(31) Geboren
(32) Zuzug aus Westdeutschland
(33) Zuzug aus dem Ausland
(38) Zuzug aus Ostdeutschland
(39) Keine Information

Alte Personen, die in einen in Vorwellen abgespaltenen Haushalt nachgezogen sind

(40) Lebt jetzt in diesem Haushalt(41-48) Analog 21-28
(99) Aktuelle Haushaltszugehörigkeit konnte in dieser Welle nicht geklärt werden.

Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment

Generally, any person appearing in the address protocol receives a code from the list above, i.e. there are no missing values for this variable.

Some notes on the different codes:
Code 0-19 apply to all households in which interviews have already taken place. Code 7 is set even when the whole household is deceased. Code 8 is set only, when a single person has moved, *not* when the entire household has moved.
Codes 20-39 apply to all persons in new households, i.e. those who lived in a household as part of a new sample, those previously interviewed founding a new household and those who moved into a newly founded household.
Codes 40-48 apply to all persons moving from an old household into a household founded earlier by another person (e.g. a child moving in with the father in wave t, where the father had moved in t-1).

$PFORM

Variable Label **„Bearbeitungsform Person“**
Value Labels $PFORM

(1) Nur Interviewer
(2) Interviewer nach positivem Telefonkontakt
(3) Schriftlich/postalisch nach positivem Telefonkontakt
(4) Nur Telefonkontakt (negativ)
(5) Telefoninterview
(6) Schriftlich/postalisch ohne Telefonkontakt
(8) Keine Bearbeitung
(9) CAPI-Interview
(10) CAWI (Computer Assisted Web Interview)

Variable format 2- digit integer
$ - Wave F10, F11, F12, F13

Comment

Note that this variable records the form of contact regardless of the result. In general, the result on the household level is identical to the one on the individual level (see $HFORMS). Hence only the last form of contact is coded.

$PERG

Variable Label **„Bearbeitungsergebnis 1-Steller“**
Value Labels $PERRG

(1) Interview Realisiert
(2) Derzeit nicht durchführbar
(3) Derzeit nicht bereit
(4) Endgültige Verweigerung
(5) Ins Ausland verzogen
(6) Verstorben
(8) Während der Feldzeit nicht auffindbar
(9) Endgültig nicht auffindbar

Variable format 1-digit integer

$ - Wave F10, F11, F12, F13

Comment $PERG shows whether an interview was conducted or not and provides basic information about possible non-response. Codes 5 and 6 mean that the person has left the sampling frame and is no longer eligible for the study. The other codes apply to the eligible (or unknown) population.

$PERGZ

Variable Label **„Bearbeitungsergebnis 2-Steller“**
Value Labels $PERRGZ

- (10) Interview Realisiert
- (11) Realisiert auch in Vorwelle
- (16) Person hat teilgenommen und Lückendaten nacherhoben
- (17) Person hat teilgenommen und Lückendaten nicht nacherhoben
- (18) Person hat teilgenommen, kein Haushaltsinterview vor
- (20) Interview derzeit nicht durchführbar:
- (21) Alt und krank
- (22) In der Schlussphase der Feldzeit nicht mehr erreicht
- (23) Ausländer: Längere Zeit im Heimatland
- (24) Krankheit/Krankenhaus über Feldende hinaus
- (25) Während der gesamten Feldphase nicht erreichbar
- (26) Geistig nicht in der Lage/Fragebogen nicht auswertbar
- (29) Sonstige unklare Fälle
- (30) Zur Teilnahme derzeit nicht bereit:
- (31) Zusage für Telefonbearbeitung, kein Fragebogen ausgefüllt
- (32) Keine Zeit/Lust, jedoch im nächsten Jahr wieder ansprechbar
- (39) Sonstige unklare Fälle
- (46) Wie Code 26, jedoch endgültig nicht mehr in der Lage
- (47) Sprachprobleme
- (50) Ins Ausland verzogen
- (60) Verstorben
- (80) Nicht auffindbar während der Feldzeit
- (90) Endgültig nicht auffindbar

Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment $PERGZ provides more detail on the interview result and allows coding more elaborate reasons for non-response.

$PADER

Variable Label **„Ergebnis der Adressenermittlung Personen“**
Value Labels $PADER

- (1) umgezogen, neue Adresse ermittelt
- (2) unbekannt verzogen
- (3) trotz Bestätigung der alten Adresse durch Post/ Einwohnermeldeamt, dort nicht auffindbar
- (4) an alter Adresse nicht auffindbar und bei Einwohnermeldeamt nicht registriert

(5) ins Ausland verzogen
(7) verzogen, wird aber nicht weiterverfolgt (Kind/ harte Verweigerer)
(8) umgezogen mit einer Person, die weiterverfolgt wird
(9) in bestehenden Panelhaushalt zurückgekehrt (laufende Personennummer ist reserviert)

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment $PADER documents the whereabouts of individuals from previously interviewed households. As most people stay in their old household, most individuals receive a "-2 Does not apply". Code 1 results in an interview request, all other depict why finding the respondent's new address was not be determined.

$PADERQ

Variable Label **„Informationsquelle der Adressenermittlung Personen"**
Value Labels $PADERQ
(1) Interviewer
(2) Post
(3) Einwohnermeldeamt
(4) Befragungsperson
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment $PADERQ documents where the information from $PADER comes from.

$AUSZUGM

Variable Label **„Monat des Auszugs einer Person"**
Value Labels $AUSZUGM (months)
Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment Shows the month in which a person left the household (the year is coded in $AUSZUGJ).

$AUSZUGJ

Variable Label **„Jahr des Auszugs einer Person"**
Value Labels $AUSZUGJ (years)
Variable format 4-digit integer
$ - Wave F10, F11, F12, F13

Comment Shows the year in which a person left the household (the month is coded in $AUSZUGM).

$EINZUGM

Variable Label	**„Monat des Einzuges einer Person“**
Value Labels	$EINZUGM (months)
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Shows the month in which a person moved into the household (the year is coded in $EINZUGJ).

$EINZUGJ

Variable Label	**„Jahr des Einzuges einer Person“**
Value Labels	$EINZUGJ (years)
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Shows the year in which a person moved into the household (the month is coded in $EINZUGM).

$ABWESM

Variable Label	**„Monat der Abwesenheit einer Person“**
Value Labels	$ABWESM (months)
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Shows the month in which a person’s temporary absence begins (see also $PZUG).

$ABWESJ

Variable Label	**„Jahr der Abwesenheit einer Person“**
Value Labels	$ABWESM (years)
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Shows the year in which a person’s temporary absence begins (see also $PZUG).

$PBIO

Variable Label **„Biographie“**

Value Labels $PBIO

(0) Kein Zusatzfragebogen nötig, da seit Welle 1984 teilgenommen
(1) Zusatzfragebogen ausgefüllt
(2) Keinen Zusatzfragebogen ausgefüllt und Grund nicht bekannt
(3) Kein Zusatzfragebogen nötig, da in akt. Welle erst ins Befragungsalter gekommen
(4) Zusatzfragebogen ausdrücklich verweigert
(5) aus 1 umgesetzt: Zusatzfragebogen vorhanden

(6) aus 2 umgesetzt: Kein Zusatzfragebogen vorhanden
(7) aus 3 umgesetzt: Kein Zusatzfragebogen nötig, da in Vorwelle/n nachgewachsen
(8) aus 4 umgesetzt: Zusatzfragebogen ausdrücklich verweigert
(9) Vor Nacherhebung des Zusatzfragebogens ausgeschieden.

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment In the SOEP, this variable denotes whether a person filled out the extra biography questionnaire or the youth questionnaire. Given that in FiD, the biography part is integrated into the p-questionnaire, this variable refers to the youth questionnaire only.

$DJ

Variable Label **„Kognitiver Test „Lust auf DJ““**
Value Labels $DJ
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable is set "-5 not asked in this sample" for all cases, as no tests were administered in FiD.

$EWSTATU

Variable Label **„Erwerbstatus von Nicht-Teilnehmern“**
Value Labels $EWSTATU
(1) In Vollzeit erwerbstätig
(2) In Teilzeit erwerbstätig
(3) Arbeitslos gemeldet
(4) In Schule/Studium/Ausbildung
(5) In Rente
(6) Sonstiges
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This Information is collected for non-responding individuals in participating households (i.e. $BEFSTAT = 3).

$EX ($E1-$E6)

Variable Label **„Elternfragebogen X“** (X=1,2,3,4,5,6)
Value Labels $EX
(1) Zusatzfragebogen ausgefüllt
(2) Keinen Zusatzfragebogen ausgefüllt und Grund nicht bekannt
(4) Zusatzfragebogen ausdrücklich verweigert
(5) aus 1 umgesetzt: Zusatzfragebogen vorhanden
(6) aus 2 umgesetzt: Kein Zusatzfragebogen vorhanden
(8) aus 4 umgesetzt: Zusatzfragebogen ausdrücklich verweigert
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment The parent questionnaires 1-6 are cohort specific questionnaires, i.e. in every wave, one questionnaire is meant to be answered by mothers (and/or fathers) of children in a specific birth year. The following table shows the distribution over the years.

Year of birth for Parent Questionnaires by Survey Year

	2010	**2011**	**2012**	**2013**
Parent Questionnaire 1	2009/2010	2011/2010	2012/2011	2012/2013
Parent Questionnaire 2	2008	2009	2010	2011
Parent Questionnaire 3	2007	2008	2009	2010
Parent Questionnaire 4	2004	2005	2006	2007
Parent Questionnaire 5	2002	2003	2004	2005
Parent Questionnaire 6	2000	2001	2002	2003

Documentation *hbrutto*

Gross information on all households

Note that some of this information is in German, based on the codebook provided by TNS-Infratest for FiD and SOEP.

List of Variables:

$HTYP

Variable Label	**„Haushaltstyp“**
Value Labels	$HTYP (1) Alter Haushalt mit unveränderter Adresse (2) Umgezogener alter Haushalt (3) Alter Haushalt, im Vorjahr umgezogen und temporärer Ausfall (4) Neuer Haushalt der Vorwelle, imVorjahr temporärer Ausfall (5) Neuhaushalt der aktuellen Welle
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
Comment	For all splits, the basic rule holds that the part remaining at the old address becomes the “old household” (code 1), whereas anyone moving will move into a “new household” (all other codes). If there are changes to codes 3, 4, or 5 within a wave, these codes remain unchanged. If the respective household ceases to exist, there is only a code 1. Code 2 is set if all members of an old household move to a new address. If a household moves AND splits at the same time, the person who was household head (or the person with the lower identification number) is defined to live in a code 2 household (and all other members receive a code 5).

$HPMAX

Variable Label	**„Höchste vergebene Personennummer im Haushalt“**
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	This variable shows, how many individual datasets have been set up in this household, i.e. how many different individuals have ever been associated with the respective household number. In case a new individual moves into the household, $HPMAX increases by one, if a household number moves out, there is no change to $HPMAX. Note that $HPMAX is very different from $HHRGR, the current household size.

$DATUMTG

Variable Label	**„Tag des letzten Haushaltskontakts“**
Value Labels	$DATUMTG (day of month)
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Refers to the day of the last contact of the household, regardless of their final state. In case the household was never in the field, the case receives code 0.

$DATUMMO

Variable Label	**„Monat des letzten Haushaltskontakts“**

Value Labels $DATUMMO (month)
Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment Refers to the month in which the household was last contacted, regardless of their final state. In case the household was never in the field, the case receives code 0.

$DATUMY

Variable Label **„Interviewjahr“**
Value Labels $DATUMY (year)
Variable format 4-digit integer
$ - Wave F10, F11, F12, F13

Comment Refers to the year in which the household was to be contacted, regardless of their final state.

$BULA

Variable Label **„Bundesland“**
Value Labels $BULA
(01) Schleswig-Holstein
(02) Hamburg
(03) Niedersachsen
(04) Bremen
(05) Nordrhein-Westfalen
(06) Hessen
(07) Rheinland-Pfalz
(08) Baden-Württemberg
(09) Bayern
(10) Saarland
(11) Berlin
(12) Brandenburg
(13) Mecklenburg-Vorpommern
(14) Sachsen
(15) Sachsen-Anhalt
(16) Thueringen
Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment $BULA refers to the federal state the household is located in during the field work. In case the household moved, the regional code is changed if necessary. In case the new address remains unknown ($HADER=2 – 4), the regional code stays as it was. Similar for the deceased ($HADER=5 – 7).

$HADER

Variable Label **„Adressermittlung Haushalt“**
Value Labels $HADER

(1) umgezogen, neue Adresse ermittelt
(2) unbekannt verzogen
(3) an alter Adresse nicht auffindbar; Post /Einwohnermeldeamt bestätigen

alte Adresse

(4) an alter Adresse nicht auffindbar; Einwohnermeldeamt hat die Person nicht registriert
(5) ins Ausland verzogen
(6) verstorben
(7) Haushaltsauflösung, Mitglieder sind in anderen Panelhaushalt verzogen
(8) Auskunftssperre gemäß Meldegesetz

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable is set for all households that are not found at their previous address. All others receive a code "-2 Does not apply". Reasons are generally a moves of address within Germany, death or moves abroad.

In case of one or more persons moving out of a remaining household, there is no code in $HADER, but in $PADER.

$HADQ

Variable Label **„Informationsquelle Adressermittlung"**
Value Labels $HADQ
(1) Interviewer
(2) Post
(3) Einwohnermeldeamt
(4) Zielperson
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment $HADQ provides the source of the information given in $HADER. It is based on the last result of the address search.

$INTZA

Variable Label **„Zahl der eingesetzten Interviewer"**
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Shows the numbers of interviewers that worked this case.

INTID

Variable Label **„Interviewer-Nummer"**
Variable format 6-digit integer
$ - Wave F10, F11, F12, F13

Comment INTID shows the unique identifier for the interviewer, which is consistently defined over all waves. In case of more than one interviewer working the household, the one conducting the household interview is the one listed in INTID. In case other interviewers did some other interviews in the same household (personal, parent), their interviewer identifier will be listed in the respective dataset.

$INTK

Variable Label **„Interviewer-Kontakte“**
Value Labels $INTNK (number of contacts)
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Provides the number of contacts until the successful interview or the refusal.

$TELK1

Variable Label **„Telefonkontakte 1“**
Value Labels $TELK1
(0) Telefonischer Bearbeitungsversuch, jedoch keine Telefonnummer zu ermitteln
(1) Telefonisch niemanden erreicht bzw. Situation nicht geklärt.
(2) Telefonisch geklärt, ob und (wenn ja) auf welche Weise weitere Bearbeitung möglich ist.
(3) Bei schriftlich-postalischer Weiterbearbeitung: Erneute telefonische Bearbeitung, da Fragebogen nach vereinbarter Frist nicht zurückgesandt wurde.
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Note that in FiD, no phone contact with the household is allowed to conduct the interview. Hence all cases are set to “-2 Does not apply”.

$TELK2

Variable Label **„Telefonkontakte 2“**
Value Labels $TELK2
(0) Keine Telefonnummer zu ermitteln
(1) Nacherhebung ohne Erfolg
(2) Nacherhebung mit Erfolg
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Note that in FiD, no phone contact with the household is allowed to conduct the interview. Hence all cases are set to “-2 Does not apply”.

$SCHK

Variable Label **„Schriftliche Kontakte“**

Value Labels $SCHK
(1) Angeschrieben ohne Reaktion
(2) Angeschrieben und Antwort erhalten
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Rarely, the household is contacted via mail, if neither the interviewer nor a phone call is successful in making contact.

$HFORM1

Variable Label **„Bearbeitungsform Haushalt“**
Value Labels $HFORM1
(1) Nur Interviewer
(2) Interviewer nach positivem Telefonkontakt
(3) Schriftlich / postalisch nach positivem Telefonkontakt
(4) Telefonkontakt mit negativem Ergebnis bzw. nicht erreicht
(5) Telefoninterview
(6) Schriftlich-postalisch ohne Telefonkontakt
(7) Mischform
(8) Keine Bearbeitung möglich
(9) CAPI
(10) CAWI (Computer Assisted Web Interview)
Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable codes the mode in which the household is first contacted to be interviewed. In case no interview is possible, but the interviewer was working the case, code 1 is set.

$HERG1

Variable Label **„Bearbeitungsergebnis Haushalt“**
Value Labels $HERG1
(0) Teilweise realisiert
(1) Vollständig realisiert
(2) Interview derzeit nicht durchführbar
(3) Haushalt derzeit zur Teilnahme nicht bereit
(4) Endgültige Verweigerung
(5) Ins Ausland verzogen
(6) Verstorben
(7) Auflösung des Haushalts wegen Verzug in anderen Panelhaushalt
(8) Während der Feldzeit nicht angetroffen / aufgefunden
(9) Haushalt endgültig nicht auffindbar
Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Documents the grade of completion in the household. Code 0 is set, if any eligible person has refused an interview request. Code 1 is only set

if all eligible persons ($BEFSTAT=1,2,3,4, or 8) were interviewed or no longer in the household ($BEFSTAT=5,6,7).

$HFORMS

Variable Label **„Schlusscode Bearbeitungsform“**
Value Labels $HFORMS

(1) Nur Interviewer
(2) Interviewer nach positivem Telefonkontakt
(3) Schriftlich / postalisch nach positivem Telefonkontakt
(4) Telefonkontakt mit negativem Ergebnis
(5) Telefoninterview
(6) Schriftlich-postalisch ohne Telefonkontakt
(7) Mischform wegen Nachbearbeitung in anderer Form
(8) Keine Bearbeitung möglich
(9) CAPI
(10) CAWI (Computer Assisted Web Interview)

Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable codes the mode in which the household is actually interviewed (as opposed to $HFORM1). In case there is only one approach to the household (as in most cases) $HFORMS and $HFORM1 are identical.

$HERGS

Variable Label **„Schlusscode Bearbeitungsergebnis“**
Value Labels $HERGS

Realisierte Haushalte
(0_) Teilweise realisiert (d.h. nicht alle Befragungspersonen haben teilgenommen)

(01) Ohne weitere Hinweise auf Folgewelle
(02) Mit Hinweis: In nächster Welle nicht mehr zur Teilnahme bereit
(03) Mit Hinweis: In nächster Welle unter keinen Umständen mehr bereit

(1_) Vollständig realisiert

(11) Ohne weitere Hinweise auf Folgewelle
(12) Mit Hinweis: In nächster Welle nicht mehr zur Teilnahme bereit
(13) Mit Hinweis: In nächster Welle unter keinen Umständen mehr bereit
(14) Vollständig realisiert ohne Personen, die BEFSTAT 7 haben
(18) Haushaltsinterview, aber kein Personeninterview vorhanden
(19) Realisierte Haushalte, die nicht gültig sind (z.B. wegen Wertgrenzen, Fälschungen)

Nicht realisierte Haushalte
(2_) Interview derzeit nicht durchführbar

(21) Alt und krank
(22) In der Schlussphase der Feldarbeit nicht mehr erreichbar
(23) Ausländer: Längere Zeit im Heimatland

(24) Krankheit / Krankenhaus über Feldende hinaus
(25) Während der gesamten Feldphase nicht erreichbar
(26) Geistig nicht in der Lage / Fragebogen nicht auswertbar
(27) Zweimal temporärer Ausfall, soll trotzdem ins Brutto der Folgewelle aufgenommen werden
(29) Sonstige / unklare Fälle

(3_) Haushalt derzeit zur Teilnahme nicht bereit

(31) Zusage zum Ausfüllen des Fragebogens in der telefonischen Bearbeitung, jedoch nicht ausgefüllt
(32) Diverse Begründungen: z.B. keine Zeit, Trauerfall; wird in Folgewelle wieder angesprochen
(39) Sonstige / unklare Fälle

(4_) bis (7_) Endgültige Ausfälle

(40) Endgültige Verweigerung
(41) Zweimal hintereinander temporärer Ausfall
(46) Endgültig nicht mehr in der Lage teilzunehmen (z.B. geistig nicht in der Lage/Pflegefall)
(47) Sprachprobleme
(48) Ganzer Point ohne Bearbeitung
(49) Einzelhaushalt ohne Bearbeitung
(50) Ins Ausland verzogen
(60) Verstorben
(70) Haushalt aufgelöst (in anderen Panel-Haushalt verzogen)

(8_) Während der Feldzeit nicht angetroffen (Adressenprobleme)

(81) Verfügt laut Auskunft Dritter über zweiten Wohnsitz, Adresse konnte jedoch nicht ermittelt werden
(82) Ständig unterwegs, auf Reisen, beruflich im Ausland
(83) Lebt laut Auskunft von Dritten überwiegend bei Freunden oder anderen Familienmitgliedern
(84) Haushalt in Adressermittlung über Feldende hinaus
(85) Neue Adresse ermittelt, jedoch nach Feldende
(89) Sonstige / unklare Fälle
(90) Haushalt endgültig nicht auffindbar
(98) QNA (Qualitätsneutraler Ausfall)
(99) Haushalt ohne Bearbeitung wegen unterschrittener Einkommensgrenze (nur Stichprobe G / Welle 2003)

Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment $HERGS shows the final result for each household. The first digit in this case is identical to the one for $HERG1, the second digit provides some more information about the code. Code 99 is never set for FiD cases.

$HSTU

Variable Label **„Bearbeitungsstufen Haushalt“**
Value Labels $HSTU

(0) Keine Bearbeitung
(1) Eine Bearbeitungsstufe
(2) Zwei Bearbeitungsstufen

(3) Drei Bearbeitungsstufen

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable documents different stages of the field work. It is related to the variable $HFORMS, which refers to the last mode of interviewing used for this household before the field period closed. Note that in FiD, most cases receive code 1, as there are no mode switches allowed.

$SPLIT

Variable Label **„Startbearbeitungsform aktuelle Welle“**
Value Labels $SPLIT

(10) Feldbearbeitung: Erfolgreiche Bearbeitung durch Interviewer in Vorwelle sowie Einzelfälle, die nicht vom Interviewer in der Vorwelle realisiert wurden (Termingründe)
(19) Eigentlich Feldbearbeitung: Erfolgreiche Bearbeitung durch Interviewer in Vorwelle, aber in aktueller Welle keine Bearbeitung wegen unterschrittener Einkommensgrenze (nur Stichprobe G 2003)
(61) Erstbearbeitung über Telefonkontakt: Temporärer Ausfall, in der Vorwelle in der telefonischen oder schriftl. Bearbeitung
(66) Erstbearbeitung über Telefonkontakt: Temporärer Ausfall, in der Vorwelle in der Interviewerbearbeitung
(68) Diverse Bearbeitungsformen: Nicht mehr im Ausgangsbrutto der aktuellen Welle enthaltene Haushalte, die aus unterschiedlichsten Gründen wieder auftauchen
(69) Erstbearbeitung über Interviewer: Ausfälle aus dem Vorjahr
(70) Erstbearbeitung über Telefonkontakt: Teilnehmer in Vorwelle, jedoch mit Hinweis auf künftige Verweigerung
(75) Keine Bearbeitung in aktueller Welle: Vor Beginn der aktuellen Welle ins Ausland verzogen
(76) Keine Bearbeitung in aktueller Welle: Vor Beginn der aktuellen Welle verstorben
(77) Keine Bearbeitung in aktueller Welle: Vor Beginn der aktuellen Welle in anderen Panelhaushalt zurückgezogen
(78) Keine Bearbeitung in aktueller Welle: In letzter Welle Teilnehmer, jedoch für Folgewellen ausdrücklich verweigert
(81) Erstbearbeitung über Telefonkontakt: Letzte Welle telefonisch/schriftlich befragt
(88) Erstbearbeitung über Telefonkontakt: Letzte Welle Telefoninterview (nur in Ausnahmefällen)
(89) Erstbearbeitung schriftlich: Letzte Welle schriftlich befragt
(90) Diverse Bearbeitungsformen: Neuhaushalte aktuelle Welle
(99) Feldbearbeitung im Erstversand: Neuhaushalte aktuelle Welle

Variable format 2-digit integer
$ - Wave F10, F11, F12, F13

Comment This variable shows the choice of the most promising form to contact the household in this wave, based on previous wave experience.

$HHGR

Variable Label	**„Zahl der im Haushalt lebenden Personen“**
Value Labels	$HHGR (number)
Variable format	2-digit integer
$ - Wave	F10, F11, F12, F13

Comment: $HHGR shows the number of people in the household at the time of the interview, not counting moved-out or deceased persons. Note that a household that has been dissolved is coded with a “0”.

$WUM1

Variable Label	**„Wohnumfeld1 - Haushaltstyp“**
Value Labels	$WUM1

(1) Landwirtschaftliches Wohngebäude
(2) Freistehendes Ein-/Zweifamilienhaus
(3) Ein-/Zweifamilienhaus als Reihenhaus oder Doppelhaus
(4) Wohnhaus mit 3 - 4 Wohnungen
(5) Wohnhaus mit 5 - 8 Wohnungen
(6) Wohnhaus mit 9 und mehr Wohnungen (aber höchstens 8 Stockwerke; also kein Hochhaus)
(7) Hochhaus, 9 und mehr Stockwerke, Wohnungen unbegrenzt
(8) Sonstiges

Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13

Comment: Note that the information for this variable is retrieved from the household questionnaire (as it is for the next two, $WUM2 and $WUM3). Only new or moved household are to answer these questions. If households do not participate, the interviewer is asked to provide the information on these variables. For old households which have stayed in the old address, there are no changes.

$WUM2

Variable Label	**„Wohnumfeld2 – Privat/Anstaltshaushalt“**
Value Labels	$WUM2

(1) Privathaushalt/Kein Wohnheim
(2) Schüler-/Jugendlichenwohnheim
(3) Studentenwohnheim
(4) Berufstätigen-/Ledigenwohnheim
(5) Altenheim
(6) Altenwohnheim
(7) Sonstiges Heim/Unterkunft
(8) Hotel / Pension

Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13

Comment: Similar to $WUM1, these data are drawn either from the household questionnaire or (in case of refusals) provided by the interviewer. This information is used to identify non-private households.

$WUM3

Variable Label	**„Wohnumfeld3 – Quartier“**
Value Labels	$WUM3 (1) Reines Wohngebiet mit überwiegend Altbauten (2) Reines Wohngebiet mit überwiegend Neubauten (3) Mischgebiet mit Wohnungen und Geschäften/Gewerbe (4) Geschäftszentrum (Läden, Banken, Verwaltungen) mit wenig Wohnungen (5) Gewerbe- und Industriegebiet mit wenig Wohnungen (6) Sonstiges
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Similar to $WUM1, these data are drawn either from the household questionnaire or (in case of refusals) provided by the interviewer.

$WEIN

Variable Label	**„Welle des Einzugs in die Aktuelle Adresse“**
Value Labels	$WEIN (year)
Variable format	4-digit integer
$ - Wave	F10, F11, F12, F13
Comment	$WEIN provides the information since when the household resides at its current address.

$HTEL

Variable Label	**„Telefon“**
Value Labels	$HTEL (1) Telefonnummer bekannt (-2) Telefonnummer unbekannt
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
Comment	Shows whether the household has a known phone number.

$INTEINS

Variable Label	**„Nummer des ersteingesetzten Interviewers“**
Variable format	1-digit integer
$ - Wave	F10, F11, F12, F13
Comment	In combination with $INTID, this variables shows whether there has been a change in interviewers during the field period.

$EMAIL

Variable Label	**„Haushalt mit e-mail Anschluss“**
Value Labels	$EMAIL

(1) Emailadresse bekannt
(-2) Emailadresse unbekannt

Variable format 1-digit integer
$ - Wave F10, F11, F12, F13

Comment Shows whether the household has a known phone number.

$MKZ1

Variable Label **„Migrantenkennzeichen1“**
Value Labels $MKZ1
(1) Adresse ohne eindeutigen Migrationshintergrund
(2) Adresse mit Migrationshintergrund laut EMA
(3) Adresse mit Migrationshintergrund laut Onomastik
Variable format 1-digit integer
$ - Wave: F10

Comment $MKZ1 defines in the cohort sample, drawn 2010, whether the household was classified as a migration household before the field period or not. This was done via nationality (code 2) and onomastic method (code 3). For the screening sample, the code is “-2 Does not apply”. This variable is only set in 2010 so far.

$MKZ2

Variable Label **„Migrantenkennzeichen2“**
Value Labels $MKZ2
(1) Deutsch
(2) Migrant/ Ausländer Basis
(3) Migrant/ Ausländer Top Up
Variable format 1-digit integer
$ - Wave: F10

Comment $MKZ2 defines in the cohort sample, drawn 2010, whether the household belonged to the Top-up sample or to the “original” migration sample. The top-up was drawn from the originally sampled households in FiD to boost the number of migrant households (see also documentation on FiD). For the screening sample, the code is “-2 Does not apply”. This variable is only set in 2010 so far.

Documentation *paradatal*[32]

Additional information regarding interview circumstances

by Mathis Fräßdorf (geb. Schröder) and Malisa Zobel

[32] Note that this file was called "$intview" up to FiD v1.2. As this name exists in the SOEP data collection but covers different information (namely information on the interviewer), this dataset was renamed to avoid confusion. The $intview dataset with the according information on the interviewers will likely be part of a future data distribution.
Up to FiDv3.0, there were wave specific file ***$paradata***. Starting with distribution 3.1, only a longitudinal dataset is included.

Contents

General Information

The data file provides information regarding:

- Questionnaire
- Interview Mode
- Duration of Interview
- Day Interview was conducted
- Month Interview was conducted
- Interviewer

The aim of the ***paradatal*** data file is to provide additional information regarding the circumstances under which the interview took place. Information is available for households, as well as individuals, based on their never-changing personal ID. With the distribution of FiDv3.1, only a longitudinal version of these datasets is available. The wave specific data can still easily be extracted by restricting to certain years using SYEAR.

Variables in paradatal

HHNR

Variable Label "**Original household number**"
Variable Format 7-digit integer

Comment HHNR is the original household number. It can be linked to the never-changing person ID (PERSNR), thereby showing the household in which the individual first entered the panel.

HHNRAKT

Variable Label "**Current wave household number**"
Variable Format 7-digit integer

PERSNR

Variable Label "**Never-changing person ID**"
Variable Format 8-digit integer

Comment PERSNR in this dataset refers to the person answering the questionnaire. Hence, in the household questionnaire, this is the household head; in the parent questionnaires, PERSNR refers to the person giving information about the child.

PERSNRK

Variable Label "**Child's never-changing person ID**"
Variable format 8-digit integer

Comment PERSNRK is the identifier complementing PERSNR in case of a parent questionnaire. It refers to the child for whom the parent questionnaire has been answered. It is set to "(-2) Does not apply" in case of the other questionnaires.

SAMPLE1

Variable label **"Subsample"**
Value label SAMPLE1
(61) FiD 2007 Birth Cohort
(62) FiD 2008 Birth Cohort
(63) FiD 2009 Birth Cohort
(64) FiD 2010 Birth Cohort
(65) FiD Screening (sampled 2010)
(66) FiD Screening (sampled 2011)
Variable format 2-digit integer

Comment Note that this variable is included in all datasets, and provides information whether the household or person originates from the cohort

or the screening sample in FiD. In ***ppfad*** and ***$kind*** it is named PSAMPLE, in ***hpfad*** it is named HSAMPLE.

QSTNR

Variable Label "**Questionnaire**"
Value Label QSTNR

(10) Parent-Qunaire 1 (0-1 yrs)
(20) Parent-Qunaire 2 (1-2 yrs)
(30) Parent-Qunaire 3 (2-3 yrs)
(40) Parent-Qunaire 4 (5-6 yrs)
(50) Parent-Qunaire 5 (7-8 yrs)
(60) Parent-Qunaire 6 (9-10 yrs)
(70) Person-Qunaire (only)
(71) Person-Qunaire (with Bio 1)
(72) Person-Qunaire (with Bio 2)
(73) Person-Qunaire (with Bio 1+2)
(80) Household-Qunaire
(90) Youth-Qunaire
(100) Luecke-Qunaire (Gap in prev. year)

Variable Format 3-digit integer

Comment QSTNR indicates which questionnaire was used during the interview. Starting with the data distribution 3.1, the codes include specific information about which person questionnaire is filled out, i.e. whether any of the biography questionnaires are included or not. (For this reason, the codes were also extended to 3-digits.)
Also, a code for the gap questionnaires is included since version 3.0. Note that even though the ***$luecke*** files collect data about the previous year (and are stored in the previous year's folder), ***paradatal*** is motivated by the surroundings of the interview (such as date or interviewer). For this reason, interview information on ***$luecke*** is kept in the year the data are collected.

REQD

Variable Label "**Questions required (not code -2,-4,-5,-6)**"
Value Label REQD
Variable Format 3-digit integer

Comment REQD counts the number of questions a respondent would have had to answer. This information varies by respondent, as some individuals may go through different parts of the questionnaire than others. E.g. employed respondents have a whole set of job related questions which individuals not in the labour market do not have to answer. This count is based on questions which are not "missing by design", i.e. "-2 does not apply", "-4 invalid multiple answers", "-5 Question not asked in sample", and "-6 Sample specific filters" are excluded.

MISS

Variable Label	"**Missing answers (code -1 or -3)**"
Value Label	MISS
Variable Format	2-digit integer
Comment	MISS provides additional information on the quality of the interview by providing the number of missing answers, i.e. counting the total of "-1 No answer" and "-3 Answer improbable" in a questionnaire.

MODE

Variable Label	"**Interview mode**"
Value Label	MODE (1) CAPI (2) PAPI (3) Mail (special)
Variable Format	1-digit integer
Comment	MODE provides information on the mode of the interview. Note that there have been mail-outs after the field phase for parent-questionnaire 2 in the screening sample of 2010. All other interviews from 2010 to 2013 have been conducted face-to-face.

PROXY

Variable Label	"**Proxy Interview**"
Value Label	PROXY (1) Condition met (2) Condition not met
Variable Format	1-digit integer
Comment	PROXY denotes that the indicated person was not able to answer the questionnaire and responses are given by a proxy. These are very few cases in FiD for the personal interviews. Note that while all parent-questionnaires are proxy interviews, we do not specifically set this variable for these questionnaires.

INTID

Variable Label	"**Interviewer Identification Number**"
Variable Format	6-digit integer
Comment	INTID provides the interviewer identification number, as given by the survey institute (TNS-Infratest). Note that there are no interviewer IDs for parent-questionnaire 2 in the screening sample of 2010, as these have been completed by mail, not with an interviewer present. These cases are coded with "(-2) Does not apply".

DURA

Variable Label "**Duration of interview**"
Variable Format 3-digit integer

Comment DURA provides information on the duration of the interview in minutes. Note that this is information provided by the interviewer, not information from the CAPI program. For all questionnaires available in PAPI (pen-and-paper) mode, the interview duration is not collected and hence set to "-2 Does not apply".

MINT

Variable Label "**Month of interview**"
Value Label MINT
(1) January
(2) February
(3) March
(4) April
(5) May
(6) June
(7) July
(8) August
(9) September
(10) October
(11) November
(12) December
Variable Format 2-digit integer

Comment MINT provides information on the month of the interview. Note that this is information provided by the interviewer, not information from the CAPI program. In case of a PAPI questionnaire (only possible in the parent questionnaires, QSTNR=1-6), the date recorded may be different from the date the respondent answered the questions.

DINT

Variable Label "**Day of interview**"
Variable Format 2-digit integer

Comment DINT provides information on the day of the month of the interview. Note that this is information provided by the interviewer, not information from the CAPI program. In case of a PAPI questionnaire (only possible in the parent questionnaires, QSTNR= 1-6), the date recorded may be different from the date the respondent answered the questions.

DOW

Variable Label "**Day of week of interview**"
Value Label DOW

(1) Monday
(2) Tuesday
(3) Wednesday
(4) Thursday
(5) Friday
(6) Saturday
(7) Sunday

Variable Format 1-digit integer

Comment DOW provides information on the day of the week the interview took place. Note that this is information provided by the interviewer, not information from the CAPI program. In case of a PAPI questionnaire (only possible in the parent questionnaires, QSTNR=1-6), the date recorded may be different from the date the respondent answered the questions.

Table of frequencies

Questionnaire	**2010**	**2011**	**2012**	**2013**
(10) Parent-Qunaire 1 (0-1 yrs)	1,321	207	212	167
(20) Parent-Qunaire 2 (1-2 yrs)	787	647	568	187
(30) Parent-Qunaire 3 (2-3 yrs)	871	741	555	523
(40) Parent-Qunaire 4 (5-6 yrs)	473	486	425	656
(50) Parent-Qunaire 5 (7-8 yrs)	682	902	849	707
(60) Parent-Qunaire 6 (9-10 yrs)	647	820	768	760
(70) Person-Qunaire (only)			5,361	6,747
(71) Person-Qunaire (with Bio 1)	7,807	1,414	143	0
(72) Person-Qunaire (with Bio 2)		6,250	1,519	0
(73) Person-Qunaire (with Bio 1+2)			154	96
(80) Household-Qunaire	4,574	4,529	4,186	3,923
(90) Youth-Qunaire	190	264	293	310
(100) Luecke-Qunaire (Gap in prev. year)			229	271
Total	**17,352**	**16,260**	**15,262**	**14,347**

Documentation *bioage17*

Detailed Information on Youths

*This documentation is based on the comparable SOEP documentation on **bioage17** and has benefited from previous work of Marco Giesselman, Mila Staneva, Henning Lohmann and Sven Witzke. For readability reasons we do not specifically cite and specify text that has been used directly from the SOEP document.*

General Information

A special group of first time respondents are young persons living in a panel household, who reach the surveying age of 17 years. From this specific group of panel entrants, we are able to obtain some more detailed information on youth and socialisation than from other new sample members. At the same time, certain life-course dimensions (as the partnership- or employment biography) have not yet developed in 17 year-olds. With regard to these specifics, the standard biography questionnaire is not appropriate to this group. Thus, we use an independent questionnaire for this special group of first time respondents: the Youth Questionnaire. This instrument was used in the SOEP since the year 2000 and was introduced in FiD from the first wave on in 2010. It can be understood as an alternative version of the Biography Questionnaire, collecting more comprehensive information on relationships with parents, leisure-time activities, and past achievements in school, as well as on personality characteristics. In addition, there are numerous prospective questions about educational plans and plans for further training, as well as questions about expectations for future career and family.

A number of statements regarding specific circumstances—including the expectations for the future mentioned above—are directly related to the time at which the questionnaire was completed. However, they provide a multifaceted background for long-term analyses since these young people will continue to be interviewed in subsequent years like other SOEP respondents. The Youth Questionnaire also contains retrospective questions, for example, at what age the teenager started his or her first job or first music lessons, what recommendations he or she received regarding choice of secondary school level, and which grades he or she repeated.

Genesis and Target Population of the Youth Questionnaire

The Youth Questionnaire is aimed at youths who have reached the surveying age of 17 years[33] and are therefore being interviewed for the first time. This questionnaire takes the place of the supplementary Biography Questionnaire, since the latter does not apply to the young people's family or career situations. As a rule, information on social origin can be obtained from the parents' Individual Questionnaire, in case the youth lives together with the respective parent. If the teenager does not live with either parent, the Youth Questionnaire collects information on the missing parent(s). Young people who immigrated to Germany are also given the standard questions on immigration from the supplementary Biography Questionnaire. This guarantees that all important information collected in the Biography Questionnaire is also available on these young people.

[33] More precisely, this refers to youths who live in an already existing panel household and are or will turn 17 years old in the year of the survey. They are therefore 16 or 17 years old at the time of the interview.

The Youth Questionnaire is used in all FiD samples, however, as Table 1 shows, the focus on young children in the Cohort Sample leads to slightly fewer youths being present in those samples. For the years currently available, 747 data on youths are provided.

Table 1: Target Population for the Youth Questionnaire by year, sample and age

Survey year	Sample			
	Cohort	Screening 2010	Screening 2011	Observations
2010	14	176	-	190
2011	20	171	73	264
2012	16	193	84	293
2013	27	205	78	310
Total	77	745	235	1,057

Contents and Structure of the Data Set bioage17

From a technical perspective, four different types of questions are asked in the Youth Questionnaire:

A) Questions used to complete certain biographical files (***bioparen***, ***bioimmig*** in the SOEP). These questions are identical to questions in the standard Biography Interview. This applies to the topic blocks 'Origin' (questions 60 to 71) and 'Childhood and Parents' House' (questions 72-85). The corresponding variables are *not* included in ***bioage17***, but combined with biographical information from non-youth new entrants in the file ***bioparen***. The information on a person's migration background are used for the variable MIGBACK in ***ppfad***, and are also available from the original dataset ***$jugend***.

B) Questions that are similar to items in the standard Biography Interview, but go further into detail. This applies to the topic blocks 'Relationships' (questions 12-14), 'Leisure and Sport' (questions 15-25) and 'Education and Career plans' (questions 26-55). These variables are stored in ***bioage17***. Corresponding Variables obtained from other new sample members (with a standard Biography Interview) are included in the original dataset ***$lela*** (in the SOEP, these data appear equivalently in ***biosoc***. Depending on the complexity and scope of the analysis, the user might want to combine corresponding data from ***bioage17*** and ***$lela*** in order to access all panel members.

C) Questions that specifically relate to young persons and therefore have no equivalent in the standard Biography Interview. This applies to the topic blocks 'Residence' (questions 1-3), 'Jobs and Money' (questions 4-11), 'Future' (question 59) and 'Attitudes and Opinions'

(questions 86-87). These Variables are stored in ***bioage17*** and have no equivalent for other panel entrants.

D) Selected time-variant questions from the regular individual questionnaire are added to the Youth Questionnaire. This refers to the questions 56 to 58, 89, 90 and the topic block 'personality' (questions 91 to 99). This data is *not* included in ***bioage17***[34], but can be accessed in the original dataset ***$jugend***.

The design of the dataset ***bioage17*** is patterned after the SOEP Youth Questionnaire, which is the standard version for subsequent years. As in the biographical data survey, every youth answers the Youth Questionnaire only once. The data is therefore presented in column form, just as it would be in a cross-sectional record. The variable ERHEBJ makes it possible to quickly identify the year of the survey.

Table 2 (at the end of this chapter) lists all variables for the dataset ***bioage17***. The first column contains the name of each variable, the second a brief specification of its content, and the third the number of the question as it appears in the Youth Questionnaire. The variables containing the identification of the person surveyed and the interview situation have no corresponding number because they do not originate from the regular section of the Youth Questionnaire.

Special Features of Some Questions and Variables

The question regarding the support received by these young people from their parents (question 14) is based on the Supportive Parenting Scale of Simons et al. (1992)[35], which was transformed for Germany by Schwarz and Walper (1997)[36]. The instrument used to compile career orientation (question 54) was taken from Kracke (1996)[37].

If the question on school attendance in the Youth Questionnaire is answered with 'yes' when at the same time information about vocational degrees is provided, a recoding is undertaken. In this case the variable BYSCHBES is changed to the value -3 (-3: Entry deleted after intensive examination).

In question 51, young people are asked whether they know what career they would like to start. If they give a positive answer ('yes, with some certainty', 'yes, with a lot of certainty'), then they are asked to specify the occupation in plain text. This plain-text entry is coded according to the classification of occupations of the Federal Statistical Office, Germany, (Statistisches Bundesamt), version 1992, and according to the ISCO 1988. In addition, the

[34] The first ten items in question 90 are still stored in BIOAGE17, for details see below.
[35] Simons, R.L., F.O. Lorenz, R.D. Conger and C.-I. Wu (1992): Support from spouse as mediator and moderator of the disruptive influence of economic strain on parenting. in: Child Development 63: 1282-1301.
[36] Schwarz, B. and S. Walper (1997): Erziehung aus Sicht von Eltern und Kindern. Erste Erfahrungen mit den Instrumenten der 1. Erhebung. Berichte aus der Arbeitsgruppe "Familienentwicklung nach der Trennung" #19/97. Ludwig-Maximilians-Universität München.
[37] Kracke, B. (1996): Fragebogen zur Berufsorientierung bei Realschülern. University of Mannheim, unpublished manuscript.

values for Ganzeboom's International Socio-Economic Index of Occupational Status (ISEI), for Treiman's Standard International Occupational Prestige Scale (SIOPS) for Erikson's and Goldthorpe's Class Category (EGP)[38] as well as Wegener's Magnitude Prestige Scale (MPS)[39] are also given.

[38] For ISCO88, SIOPS, ISEI and EGP see Ganzeboom, H.B.G. and D.J. Treiman (1996): Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. in: Social Science Research 25, 201-239.

[39] Frietsch, R. and H. Wirth (2001): Die Übertragung der Magnitude-Prestigeskala von Wegener auf die Klassifizierung der Berufe. in: ZUMA-Nachrichten, 48, 139-163.

Table 2: Description of the data set BIOAGE17

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
	Entries for surveyed person	
HHNR	Original household identifier (invariant)	
HHNRAKT	Actual household identifier	
PERSNR	Personal identifier	
BEFRPER	Respondent identifier	
ERHEBJ	Survey year	
BYGEBJAH	Year of birth	
BYMNR	identifier of mother (taken form BIOPAREN; social, not necessarily biological relationship)	
BYVNR	identifier of father (taken from BIOPAREN; social, not necessarily biological relationship)	
	Residence	
BYWOELT	Residing in parents` household (HH)	01
BYWOZIM	Own room	02
BYWOWEI	Additional apartment outside of parents' HH	03
	Jobs and Money	
BYVDEIG	Own income	04
BYVDART	Type of income	05
BYJBFRUE	Worked before (on holiday or while in school)	06
BYJBALT	Age by first job (on holiday or while in school)	07
BYJBGRUN	Reason for working	08
BYTGELD	Allowance	09
BYTGELDW	Amount of allowance per week	10
BYTGELDM	Amount of allowance per month	10
BYSPAR	Saving money	11
BYSPARM	Amount saved every month	11
BYSPARUN	Sporadic saving	11
	Relationships	
	Importance of various persons:	
BYWIVA	Father	12
BYWIMU	Mother	12
BYWIBS	Brother, Sister	12
BYWIVW	Other related persons	12
BYWIFFR	Serious boy/girlfriend	12
BYWIBFR	Best friend	12
BYWILEHR	Teacher	12
BYWICLQ	Clique	12
BYWISON	Other person	12
	Frequency of fights with:	
BYSTRVA	Father	13
BYSTRMU	Mother	13
BYSTRBS	Brother, Sister	13
BYSTRFFR	Serious boy/girlfriend	13
BYSTRBFR	Best friend	13
BYBZ01MU	Talk with mother about personal experiences	14
BYBZ01VA	Talk with father about personal experiences	14
BYBZ02MU	Mother addresses problems	14
BYBZ02VA	Father addresses problems	14
BYBZ03MU	Mother asks opinion before a decision is made	14
BYBZ03VA	Father asks opinion before a decision is made	14

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
BYBZ04MU	Mother shows approval	14
BYBZ04VA	Father shows approval	14
BYBZ05MU	Solve problems together with mother	14
BYBZ05VA	Solve problems together with father	14
BYBZ06MU	Mother shows trust	14
BYBZ06VA	Father shows trust	14
BYBZ07MU	Mother asks opinion on family issues	14
BYBZ07VA	Father asks opinion on family issues	14
BYBZ08MU	Mother justifies decision	14
BYBZ08VA	Father justifies decision	14
BYBZ09MU	Mother shows love	14
BYBZ09VA	Father shows love	14
	Leisure and Sports	
	Frequency of free time activities:	
BYFZFERN	TV, Video	15
BYFZPC	Computer games	15
BYFZMUSH	Listen to music	15
BYFZMUSS	Play music	15
BYFZSPRT	Do sports	15
BYFZTANZ	Dance, Theatre	15
BYFZTECH	Technical work, Programming	15
BYFZLESE	Read	15
BYFZEHRE	Volunteer activities	15
BYFZABH	Do nothing, hang around, day dream	15
BYFZMFFR	Spend time with boy/girlfriend	15
BYFZMBFR	Spend time with best friend	15
BYFZMCLQ	Spend time with clique	15
BYFZINT	Internet/chatting	15
BYFZJUGZ	visiting youth center	15
BYFZRELI	go to church/religious activities	15
BYMUSSP	Actively make music	16
BYMUSART	Style of music made	17
BYMUSMW	Play music with whom	18
BYMUSALT	Age starting playing music	19
BYMUSUNT	Paid music lessons	20
BYSPRTTR	Participate in sports	21
BYSPRTAR	Favourite sport	22
BYSPRTAL	Age started favourite sport	23
BYSPRTMW	Where and with whom favourite sport	24
BYSPRTWE	Participation in competitions	25

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
	Education and Career Plans	
BYSCHBES	School attendance	26
BYSCHEND	Last year of school	27
BYSCHABS	Type of school certificate	28
BYSCHZUK	Strive for further school certificate	29
BYSCHZAR	Type of further school certificate	30
BYFMD1	1. foreign language	31
BYFMD2	2. foreign language	31
BYSCHAUS	School attendance in foreign country	32
BYSCHPRI	Attendance in a private school	33
	Activities in school:	
BYENKSPR	Class representative	34
BYENSSPR	School representative	34
BYENSZTG	School newspaper	34
BYENTHEA	Theatre, Dance group	34
BYENCHOR	Choir, Music	34
BYENSPRT	Sport group	34
BYENSONS	Other groups	34
BYENNEIN	No activities	34
BYZFINSG	Satisfaction with effort at school (overall)	35
BYZFDEUT	Satisfaction with effort in German	35
BYZFMATH	Satisfaction with effort in math	35
BYZFFMD1	Satisfaction with effort in 1. foreign language	35
BYEMPFEH	Recommendation after elementary school	36
BYNTDEUT	Last grade[40] in German	37
BYNTMATH	Last grade in math	37
BYNTFMD1	Last grade in 1. foreign language	37
BYPTDEUT	Total points[41] in German	37
BYPTMATH	Total points in math	37
BYPTFMD1	Total points in 1. foreign language	37
BYGSDEUT	Level of German at comprehensive school[42]	37
BYGSMATH	Level of math at comprehensive school	37
BYLKDEUT	Complementary / main subject[43] in German	37
BYLKMATH	Complementary / main subject in math	37
BYLKFMD1	Complementary / main subject in 1. foreign language	37
BYKLWDJA	Class repeated	38
BYKLWD1	Class level 1. repeated	39
BYKLWD2	Class level 2. repeated	39
BYNACHHI	Paid tutor lessons	40

[40] Students normally receive grades ranging from 1 to 6, whereby 1 is the best and 6 the worst. This system of assigning grades is used up to the 11th or 12th grade (level II of upper secondary or comprehensive school) depending on the federal state. After that, a new grading system is used. To make the data set more user-friendly, the information given for school grades and the information on points transformed into grades is stored in this variable. Note: No corrections have been made when a person has reported both grades and point scores and when the two types of information do not correctly correspond.

[41] From the 11th or 12th grade on, pupils are awarded points in upper secondary or comprehensive school ranging from 0 to 15, whereby 15 points are the best, 0 points the worst. The link between points and grades is as follows: 0 points: 6; grade of 1 to 3 points: grade of 5; 4 to 6 points: grade of 4; 7 to 9 points: grade of 3; 10 to 12 points: grade of 2; 13 to 15 points: grade of 1.

[42] The subjects German, math and the first foreign language are split up into different levels during the secondary school level I in comprehensive schools. Level A is the highest. The number of levels differs between the federal states.

[43] From the 11th or 12th grade on, pupils can choose their main subjects. At this stage, German, math and foreign languages can be downgraded from major to minor subjects.

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
BYELKUEM	Parents care about efforts at school	41
BYELHAUS	Parents help with homework	42
BYELDIFF	Problems with parents because of effort at school	43
BYELABEN	Parents attend parents' evening	44
BYELSPRE	Parents go to parents' day	44
BYELLEHR	Parents go to see a teacher	44
BYELVERT	Active as parent representative	44
BYELNIDA	Parents do not participate in any of these activities	44
BYKLAUSL	Number of foreign classmates	45
BYBAABGE	Vocational education, Internship, training	46
BYBABGJ	Vocational introductory year ("Berufsgrundschul- / Berufsvorbereitungsjahr")	47
BYBABEGL	Vocational integration training ("Berufl. Eingliederungslehrgaenge")	47
BYBALEH	Vocational education, apprenticeship ("Berufsausbildung, Lehre")	47
BYBABFS	Full-time vocational school/ School for public health ("Berufsfachschule / Schule des Gesundheitswesen")	47
BYBAPRAK	Internship ("Praktikum, Voluntariat")	47
BYZAJA	Vocational / university degree is aspired	48
	Type of aspired vocational / university degree:	
BYZALEH	Apprenticeship ("Lehre")	49
BYZABFS	Full-time vocational school/ School for public health ("Berufsfachsule / Schule des Gesundheitswesen")	49
BYZAFSC	Technical school, school for master of a trade ("Fachschule, Meister-, Technikerschule")	49
BYZABEA	Training for civil servants (officer) ("Beamtenausbildung")	49
BYZABAK	Approved vocational academy ("anerkannte Berufsakademie")	49
BYZAFH	Advanced technical college ("Fachhochschule")	49
BYZAUNI	University	49
BYSLBALT	Desired age for financial independence	50
BYBWUNJA	Occupation is aspired	51
	Occupation categories, encoded:	
BYKLAS	Classification of career according to the Federal Statistical Office, Germany, (Statistisches Bundesamt), version 1992	52
BYISCO88	International Standard Classification of Occupation 1988 (ISCO88)	52
BYEGP	Erikson and Goldthorpe's Class Category (EGP)	52
BYISEI	International Socio-Economic Index of Occupational Status after Ganzeboom (ISEI)	52
BYSIOPS	Treiman's Standard International Occupational Prestige Scale (SIOPS)	52
BYMPS	Magnitude Prestige Scale after Wegener (MPS)	52
BYZBINF	Information level of planned career	53
BYZBELT	Influence of the parents on career choice	54
BYZBLAS	No specific career in mind	54
BYZBBES	Intensive thoughts about various careers	54
BYZBRAU	Still looking for a career	54
	Important aspects for the career choice:	
BYWBSICH	Secure job	55
BYWBEINK	High income	55

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
BYWBAUF	Promotion opportunities	55
BYWBANE	Established profession	55
BYWBFREI	Enough free time	55
BYWBINT	Interesting activities	55
BYWBSELB	Working independently	55
BYWBKONT	Contact with persons	55
BYWBGSL	Relevant to society	55
BYWBGSND	Healthy conditions at work	55
BYWBFAM	Flexibility for family	55
BYWBHELF	Help others	55
	Future	
	Probability of future career related and private events:	
BYWAAUSP	To be accepted for a desired apprenticeship / place at university	59
BYWAERFA	To complete training/ university successfully	59
BYWAARBP	Job in desired career	59
BYWABERF	Job-related success	59
BYWAARBL	Longer unemployment	59
BYWAZURU	From family related reasons held back in career	59
BYWASELB	Self-employed	59
BYWAAUSL	Work in foreign country	59
BYWAHEIR	To marry	59
BYWAPART	Live together with partner (not married)	59
BYWAKID1	Have one child	59
BYWAKIDM	Have two or more children	59
	Attitudes and Opinions	
BYGLPART	Happiness: live with/without partner	86[44]
BYGLKIND	Happiness: with/without children	87[45]
	Success in FRG from	
BYEFFLEI	Studiousness	88[46]
BYEFAUSN	Exploitation of others	88
BYEFINT	Intelligence	88
BYEFFAM	Family's origin	88
BYEFFACH	Technical know-how	88
BYEFGELD	Money	88
BYEFSABS	School education	88
BYEFHART	Being inconsiderate and hard	88
BYEFBEZ	Networking	88
BYEFPOLI	Political activities	88
BYEFMANN	Sex/ 'being a man'	88
BYEFINI	Being dynamic and taking initiative	88
BYESVERL	What happens in life, depends on me	90[47]
BYESERRE	Did not reach, what I deserve	90
BYESGLUE	What you achieve, is a matter of luck	90
BYESAND	Others decide about my life	90
BYESHART	You have to work hard for success	90
BYESZWEI	By difficulties, doubt about own abilities	90
BYESSOZU	Chances are determined by social circumstances	90

[44] Question 88 in 2012/2013
[45] Question 89 in 2012/2013
[46] Question 90 in 2012/2013
[47] Question 92 in 2012/2013

Variable Name	Content of the Variable	Number of Question in Youth Questionnaire
BYESFAEH	Abilities are more important than efforts	90
BYESKNTR	Little control over events in my life	90
BYESENGA	Change of social circumstances through social/political activities	90
	Specification of Interview Situation	
BYINTA	Type of interview	
BYDAUER1	Duration of personal interview	
BYDAUER2	Duration of interview filled out independently	
BYANW	Presence of other persons	
BYTAGIN	Day of the interview	
BYMONIN	Month of the interview	
INTID	Identifier of the interviewer	

Documentation *$bioparen*

Biography Information for the Parents of FID-Respondents

by Linda Wittbrodt

*(This documentation is based on the SOEP versions of the **bioparen** documentation and has benefited from work by Anne Fromm, Sebastian Frischholz, Daniel D. Schnitzlein, Charlotte Büchner, Stefanie Lenuweit, Katharina Mahne, Matthias Pollmann-Schult, Jürgen Schupp and Verena Tobsch. Please understand that for readability reasons we do not mark text that has been directly taken from earlier work.)*

Short summary

The aim of the data file ***bioparen*** is to make the biography entries on the parents and on the social origin of the respondent available.

How biography information has been collected in the FID

Since the first wave of FID in 2010 the respondents received a separate Biography Questionnaire ('Lebenslauf-Fragebogen') in addition to the Individual Questionnaire. In 2010 it included only questions about the history of the surveyed persons themselves. In 2011 (the second wave of FID) another part of the Biography Questionnaire was handed out, where intergenerational aspects of the persons surveyed were included by means of a special group of questions. This deals with statements made about the education or professional training of the parents, the parents' residency, and their year of birth and death. In 2012 and 2013 the complete collection of biography questions was included in only one Biography Questionnaire for individuals surveyed for the first time. However, those who had only answered either of the two separate parts handed out in wave one and two were given the other, still unanswered part.

Therefore there are persons who filled in the two Biography Questionnaire parts separately as well as persons who were given the complete questionnaire. In addition there are respondents who only answered one of the two parts because they dropped out of the study at some point.

This is the reason why, in contrast to the SOEP, there are persons who answer two Biography Questionnaires (but different parts of it) in two different years.

In addition to the Biography Questionnaire, there is an independent questionnaire (Youth Questionnaire) in FID for the group of survey participants who are 17 years old and are being interviewed for the first time; this questionnaire is mostly identical to the Biography Questionnaire.

How is bioparen generated?

The information available in ***bioparen*** is obtained in two different ways. On the one hand, ***bioparen*** includes the children's proxy entries on the parents from the Biography Questionnaire and the Youth Questionnaire. On the other hand, it contains the direct entries from the parents in the case the respondent lives in the same household as his parents.

Every respondent is asked for information on the regional mobility of the children, adolescents also on the religious affiliation of the parents. However, information on the year

of birth, as well as the education and occupational training of the parent, additional to the professional position and occupation of father or mother are, due to the filter command in the questionnaire, not collected when the parent lives in the same household as the child at the time of the survey. In this case, the direct entries of the parents are used.

The identification of the parents occurs first of all through the variable $STELL (relationship to head of household). The possible values of the variable $STELL are listed in Table 2. The combinations of these characteristics of the $STELL-variable and their assigned interpretation for the generation of parent identifiers are describe in Table 3.

The second source of information is the population of the file ***$kind***, which includes all children under the age of 16. The file contains the personal number of the mother, as well as the personal number of the father. Through both variables the latest mother, as well as the father are identified, ideally, at the time when the child is 16 years old and thus one year before the first survey of the child. In the case the parents could not be identified by the $STELL variable, this information is used.

In a further step the biological mother or father are identified through the parent-child relationship in the file ***biobirth***. In the event that still no personal number for the mother or the father exists, the number from ***biobirth*** is used.

Table 1: Number of observations in BIOPAREN

Year of data collection	Sample						**Total**
	61	62	63	64	65	66	
2010	165	188	160	171	1.046	0	1.730
2011	757	778	736	769	3.282	280	6.602
2012	56	55	45	43	441	1.321	1.961
2013	13	10	12	9	261	94	399
Total	991	1.031	953	992	5.030	1.695	10.692

Table 2: Characteristics of the variable $STELL "relationship of the person to the head of the household"[48]

Code	Description
0	HH
1	Marital partner of the HH
2	partner of the HH
3	Daughter/son (also adopted/stepchild) of the HH
4	Foster child of the HH
5	Daughter/son–in-law of the HH
6	Father/mother of the HH
7	Father/mother-in-law of the HH
8	Brother/sister, brother/sister-in-law of the HH
9	Grandchild of the HH
10	Other relationship to the HH
11	Not related to the HH
12	Daughter/son of the partner of the HH
13	Marital partner of the HH (same sex)

Table 3: Possible Parent-Child Relationships based on $STELL

Relationship of the child to the HH	Relationship of the parent to the HH	Person is ...
3	0	Child of HH
3	1 or 2	Child of marital/ partner of HH
4	0	Foster child of HH
4	1 or 2	Foster child of marital/ partner of HH
12	2 or 3	Child of partner of HH
9	3 or 4	Child of child/foster child of HH
0	6	Child is HH, lives with parents in same household
1 or 2	7	Marital partner/partner of HH (child of in laws of HH)
9	5	Grandchild of HH (child of son/daughter-in-law of HH)

[48] Starting in F12, $STELL was changed and includes more values than in previous waves, which allows more precise identifications of relationships within the households. For details on the conversion of $STELL from 2012 to the Version of 2010/2011 see the documentation on ***$kind***.

List of Variables:

VNR / MNR

Variable Label: Personal number of the father of the respondent / Personal number of the mother of the respondent
Variable Format: 8-digit integer

Values:
(-1) PERSNR father / mother unknown
(-2) Does not apply
(-3) Answer improbable

Description: The personal ID of the parents (VNR and MNR) is generated in three steps.
1. The parents of the respondent are identified by the relationship to the head of the household ($STELL in ***$brutto***). Ideally, the children's parents are identified at the time of the first survey of the child, i.e., when the child is 17 years old. Furthermore, the social parents and not necessarily the biological parents are identified.
2. The parents of the respondent are identified via the mother's and father's ID in ***$kind***. By using these variables the "oldest" parents are identified. Ideally, these are the parents at the time the child is 16 years old (one year before the first survey).
3. The mother-ID as well as the father-ID of the respondent can be identified in ***biobirth***.
As ***bioparen*** aims at identifying the parents that live in the household when the child is 17 years old, the steps above are carried out in the hierarchy 1-3 with step 1 having the highest priority. If one is interested in only biological parents, please have a look at the information in ***biobirth***.

VGEBJ / MGEBJ

Variable Label: Year of birth of the father / Year of birth of the mother
Variable Format: 4-digit integer

Description: In a first step the information of the year of birth comes from the Biography Questionnaire. Due to a filter command, the children's proxy entries are only available for these variables when the parents or one parent and the child do not live in the same household at the time of the survey.
After the parents' personal numbers have been identified the information can be compared with the entries in PPFAD. If there are differences of +/- two years the VNR / MNR will be set to missing. The same applies when parents were aged less than 10 at the time of birth.
For the missing entries the information of the parents' year of birth is taken from PPFAD.

VTODJ / MTODJ

Variable Label: Year of death of the father / Year of death of the mother
Variable Format: 4-digit integer

Description: The variables are generated as usual using the information from the Youth Questionnaire or the Biography Questionnaire and the parents' direct-entries from PPFAD. As a next step the annual proxy information on a parent's death

from the $P-files are used. Furthermore we use information of the month of death of a parent from the year before. With this data a wrong marking as “no death in 2010” / “death in 2011” can be corrected if there is data from 2011 indicating that one parent died e.g. in October 2010.
The variables VTODJ and MTODJ will be updated with new survey information. They are updated as long as the father or the mother is part of the FID sample. We additionally use the annual proxy information of respondents about reported life events of the last year.

VAORT11 / MAORT11

Variable Label Residency of the father/ Residency of the mother in 2011
Variable Format 2-digit integer
Values

(0) Has Died	(6) Lives Elsewhere In Germany
(1) Lives In Same HH	(7) Lives Elsewhere
(2) Lives In Same Housing	(8) Lives Else E Germany
(3) Lives Neighborhood	(9) Lives Else W Germany
(4) Lives Same Town	(10) Lives Foreign Country
(5) Lives Other Town	

Description The information on the residency of the parents stems from the Youth and Biography Questionnaires as well as from the Person Questionnaire.
The information from the $P-files have a higher priority. For more details see also the information on VAORTAKT / MAORTAKT.

VAORTAKT / MAORTAKT

Variable Label Father‘s place of residence /Mother‘s place of residence
Variable Format 2-digit integer
Values

(0) Has Died	(6) Lives Elsewhere In Germany
(1) Lives In Same HH	(7) Lives Elsewhere
(2) Lives In Same Housing	(8) Lives Elsewhere E Germany
(3) Lives Neighborhood	(9) Lives Elsewhere W Germany
(4) Lives Same Town	(10) Lives Foreign Country
(5) Lives Other Town	

Description The variables VAORTAKT and MAORTAKT contain the latest available information about the parents’ residence and on whether or not they are deceased, respectively.
For persons without identified parents who answered the biography questionnaire in 2011, the information from the Person Questionnaire in 2011 was assumed.
For those persons whose parents are identified in the FID, the information on the year of death in PPFAD was used for updating. If the year of death lies

chronologically after the latest available information, VAORTAKT and MAORTAKT were put on "deceased".[49]

VAORTUP / MAORTUP

Variable Label Year of update of VAORTAKT/MAORTAKT
Variable Format 4-digit integer

Description The variable contains the year, in which the information stored in VAORTAKT and MAORTAKT has been updated.

VSBIL / MSBIL

Variable Label Education of the father / Education of the mother
Variable Format 1-digit integer

Values

(0) Do Not Know
(1) Secondary School Degree
(2) Intermediate School Degree
(3) Technical School Degree
(4) Upper Secondary School Degree
(5) Other Degree
(6) No School Degree
(7) School Not Attended

Description The parents' education is generated with information from the Youth Questionnaire, the Biography Questionnaire and direct entries from the ***$pgen***-files. Due to the filter command, the children's proxy entries are only available for VSBIL / MSBIL when the parents or one parent and the child do not live in the same household at the time of the survey.

[49] In gathering the information from different data sets, inconsistencies occurred. On the one hand, some parents had been reported as deceased in the early waves, while information about their residence at a later date was available. In this case, the information about the parents' residence was not accepted.

VBBIL / MBBIL

Variable Label Vocational training of the father / Vocational training of the mother
Variable Format 2-digit integer

Values

(0) Do Not Know
(10) No Vocational Degree
(20) Vocational Degree
(21) Trained in Foreign Company
(22) Trained long Time in Foreign Company
(23) Foreign Vocational School
(24) Trade, Farming Apprentice
(25) Business Apprentice
(26) Health Care School
(27) Special Technical School
(28) Civil Service Training
(30) Tech Engineer School
(31) Foreign Collage
(32) College, University
(40) Other Training
(50) Currently in Vocational Training
(51) Currently in Schooling

Description The parents' vocational training is generated the same way as the education variables (see VSBIL / MSBIL).

VSINFO / MSINFO

Variable Label Origin of the information on father's education /
Origin of the information on mother's education
Variable Format 1-digit integer

Values:

(0) Do Not Know
(1) Biography-Proxy
(2) $P-Individual Info

Description: The variable contains the origin of the information on parental education.

VBINFO / MBINFO

Variable Label Origin of the information on father's vocational training /
Origin of the information on mother's vocational training
Variable Format 1-digit integer

Values

(0) Do Not Know
(1) Biography-Proxy
(2) $P-Individual Info

Description The variable contains the origin of the information on parental vocational training.

VRELI / MRELI

Variable Label Religious affiliation of the father / Religious affiliation of the mother
Variable Format 1-digit integer

Values

(0) Do Not Know – Proxy
(1) Catholic
(2) Protestant
(3) Other Christian Denomination
(4) Islamic Denomination
(5) Other Denomination
(6) No Denomination

Description The questions about the religious affiliation of the parents are only asked to youngsters who are not living in the household of their parents.

VNAT / MNAT

Variable Label Nationality of the father / Nationality of the mother
Variable Format 1-digit integer

Values

(1) German
(2) Other

Description The information on the parents' nationality is generated similar to VRELI / MRELI. The question is only asked to youngsters who are not living in the same household as their parents. In addition, the parents' personal numbers are used to match information on parents' nationality with data from the $PGEN-files in case there are missing entries.

VBSTELL / MBSTELL

Variable Label Professional position of the father (when the respondent was 15 years old) / Professional position of the mother (when the respondent was 15 years old)
Variable Format 3-digit integer

Description The children's proxy entries on professional position and occupation of the father or mother (VBSTELL / MBSTELL) as well as VISCO88 /MISCO88 and all prestige scores are available when the parent and the child do not live in the same household at the time of the survey and if the parent lived in Germany when the child was 16 years old. Besides the proxy entries parents' direct information from the $P-files are used.

VBSINFO / MBSINFO

Variable Label: Origin of the information on the professional position of the father / Origin of the information on the professional position of the mother
Variable Format: 1-digit integer

Values
(0) do Not Know-Proxy
(1) Biography-Proxy
(2) $P-Individual Info

Description: The variables VBSINFO / MBSINFO are indicator variables. They tell whether the information is from the Biography or Youth or Person Questionnaires. This information is generated at the same steps as it is done with the VBSTELL / MBSTELL variables.

VISCO88 / MISCO88

Variable Label: Professional occupation of the father (when the respondent was 15 years old) / Professional occupation of the mother (when the respondent was 15 years old)
Variable Format: 4-digit integer

Description: The variables contain the ISCO88 code for the father and mother. See also documentation on ***$pgen*** (variable IS88$$).

VISEI / MISEI

Variable Label: Prestige score of father – concept of Ganzeboom / Prestige score of mother – concept of Ganzeboom
Variable Format: 2-digit integer

Description: The variables contain the ISEI code for the father and mother. See also documentation on ***$pgen*** (variable ISEI$$).

VMPS / MMPS

Variable Label: Prestige score of father – Magnitude scale – Wegener / Prestige score of mother – Magnitude scale – Wegener
Variable Format: 5-digit real

Description: The variables contain the prestige scores (magnitude scale - Wegener) for the father and mother. See also documentation on ***$pgen*** (variable MPS$$).

VSIOPS / MSIOPS

Variable Label: Prestige score of father – Treiman standard score / Prestige score of mother – Treiman standard score
Variable Format: 2-digit integer

Description The variables contain the prestige scores (Treiman standard score) for the father and mother. See also documentation on ***$pgen*** (variable SIOPS$$).

VEGP / MEGP

Variable Label Prestige score of father – Erikson – Goldthorpe class category/
Prestige score of mother – Erikson – Goldthorpe class category
Variable Format 2-digit integer
Values (-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description The variables contain the prestige scores (EGP) for the father and mother. See also documentation on ***$pgen*** (variable EGP$$).

VBKLAS / MBKLAS

Variable Label Occupational coding scheme father according German statistical office/
Occupational coding scheme mother according German statistical office
Variable Format 4-digit integer

Description: The variables contain the occupational code for the father and mother according to the coding scheme of the German statistical office. See also documentation on ***$pgen*** (variable CLASS$$).

ORTKINDH / ORTKIND1

Variable Label ORTKINDH Place of childhood /
ORTKIND1 Still lives in place of childhood?
Variable Format 1-digit integer
Values

ORTKINDH:	**ORTKIND1:**
(1) Large City	(1) Yes, Still
(2) Medium City	(2) Yes, Again
(3) Small City	(3) No
(4) Countryside	

Description The variables provide information on the place of childhood.

LIVING1 - LIVING8

Variable Label	LIVING1	No. of years living with both parents
	LIVING2	No. of years living alone with mother
	LIVING3	No. of years living with mother and new partner of mother
	LIVING4	No. of years living alone with father
	LIVING5	No. of years living alone with father and new partner of father
	LIVING6	No. of years living with other relatives
	LIVING7	No. of years living with foster parents
	LIVING8	No. of years living in youth center

Variable Format 2-digit integer
Description The variables show the total number of years for different categories of where the child lived during his childhood.

VSTREIT / MSTREIT

Variable Label Conflict with father /
Conflict with mother
Variable Format 1-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(1) Very Often
(2) Often
(3) Sometimes
(4) Seldom
(5) Never
(6) Person Not Present

Description The variables provide information on the frequency of conflicts with the parents. It is only asked in the Youth Questionnaire.

BIOYEAR

Variable Label Year of the Biography Survey
Variable Format 4-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description The variable BIOYEAR provides the year in which the information was surveyed. Since in FiD, the Biography Questionnaire is split up into two parts for individuals first interviewed in 2010 and 2011, respondents generally answered the biography in two different years. BIOYEAR then provides the year when the second part (in which the main part about a respondent's parents is collected) is answered. If the second part was never answered, the year of filling out the first biography part is stored in BIOYEAR. For first time respondents of 2012 or later, the biography is not split up. Youths do not fill out a biography questionnaire, so the year of filling out the youth questionnaire appears in BIOYEAR. For all cases, BIOYEAR is changed to the most recent year in which any sort of information was taken into the data.

BIO

Variable Label Form of Biography Questionnaire
Variable Format 1-digit integer

Values
(-1) No answer
(-2) Does not apply

(-3) Answer improbable
(1) Youth
(2) Biolela blue
(3) Only first part of Biography Questionnaire
(4) Only second part of Biography Questionnaire
(5) Both parts of Biography Questionnaire

Description The variable BIO is generated to indicate the origin of the information in BIOPAREN (Youth Questionnaire or one or both parts of the Biography Questionnaire).

ALTER / VALTER / MALTER

Variable Label Age of the respondents / Age of the respondent´s father / Age of the respondent`s mother
Variable Format 2-digit integer

Values (-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description The variable ALTER gives the age of the respondent at the moment of the interview. VALTER gives the age of the respondents' father when the respondent answered the Biography Questionnaire or the Youth Questionnaire. The same was applied for the mothers with the variable MALTER. In order to generate the variables the information for the parents who are identified in the FID was gained with data from PPFAD. The proxy entries from BIOPAREN were used when there weren't any information of the respondents parents available.

VORIGIN / MORIGIN

Variable Label Country of origin of the respondent's father / Country of origin of the respondent's mother
Variable Format 3-digit integer

Values (-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description These variables give information about the country of origin of the respondents mother (MORIGIN) and father (VORIGIN). This information is collected in the Youth and the Biography Questionnaire. Another source of information can be found in PPFAD by the direct-entries of the parents in the variable CORIGIN. These two kinds of information, proxy- and direct-entries, are used to generate MORIGIN and VORIGIN. In a first step we use the proxy-information for all the parents whose children made an entry in the Youth Questionnaire. For all the parents where there are no proxy-information available, we then use the direct-entries of the parents from the PPFAD-variable CORIGIN.

GESCHW

Variable Label Siblings yes/no
Variable Format 1-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(1) Yes
(2) No

Description GESCHW contains the information whether a respondent has siblings or not. The question was asked in the Youth Questionnaire in 2010 and since 2011 it is asked in both Youth and Biography Questionnaire.

GESCHWUP

Variable Label Time of update – siblings
Variable Format 4-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description: GESCHWUP contains the year, in which the latest sibling information was surveyed.

NUMS

Variable Label Number of sisters
Variable Format 2-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description: NUMS contains the number of sisters. The question was asked in the Youth Questionnaire in 2010 and since 2011 it is asked in both Youth and Biography Questionnaire.

NUMB

Variable Label Number of brothers
Variable Format 2-digit integer

Values
(-1) No answer
(-2) Does not apply
(-3) Answer improbable

Description NUMB contains the number of brothers. The question was asked in the Youth Questionnaire in 2010 and since 2011 it is asked in both Youth and Biography Questionnaire.

TWIN

Variable Label	Twin sister/brother
Variable Format	1-digit integer

Values:
(-1) No answer
(-2) Does not apply
(-3) Answer improbable
(1) Yes, monozygotic
(2) Yes, dizygotic
(3) No

Description: TWIN contains information whether the respondent has a twin sibling. The question was asked in the Youth Questionnaire in 2010 and 2011

Änderungen in den Versionen 1.1-4.0 der FiD Datenweitergaben

Nummer	Datensatz	Beschreibung	in Version
1.1.1	bioage03	Die Variablen "b03stcaremy" und "b03stcareyry" bezeichneten nicht den Monat und das Jahr, in dem die Betreuung angefangen wurde.	**1.1**
1.1.2	bioage03	Die Variablen "b03spe6" und "b03spe5" sind bei der Screening-Stichprobe miteinander vertauscht worden.	1.1
1.1.3	bioage03	Die Labels der Variablen "b03spe7", "b03spe8b", "b03spe6", und "b03spe5" waren vertauscht	1.1
1.1.4	bioage03	Die Variablen "b03health8", "b03health1" und "b03health7" sind in der Screening-Stichprobe miteinander vertauscht worden.	1.1
1.1.5	bioage10p1/p2	Die Variablen "b10imper1e"-"b10imper3e" wurden neu gelabeled.	1.1
1.1.6	bioage10p1/p2	Die Variablen "b10biopar" und "b10sexresp" waren nicht aufeinander abgestimmt.	1.1
1.1.7	f10eltern1	Geburtsmonate wurden verändert.	1.1
1.1.8	f10eltern2	Geburtsmonate wurden verändert.	1.1
1.1.9	f10eltern3	Geburtsmonate wurden verändert.	1.1
1.1.10	f10eltern3	Die Variablen "f10e337e", "f10e337f" und "f10e337g" sind in der Screening-Stichprobe miteinander vertauscht worden.	1.1
1.1.11	f10eltern3	Die Variable "f10e335a6" und "f10e335a7" sind bei der Screening-Stichprobe miteinander vertauscht worden.	1.1
1.1.12	f10eltern4	Durch falsche Bespaltung sind die Variablen "f10e404b" (Größe) und "f10e404a" (Gewicht) nicht korrekt eingelesen worden.	1.1
1.1.13	f10eltern5	Geburtsmonate wurden verändert.	1.1
1.1.14	f10eltern5	Geburtsmonate wurden verändert.	1.1

Nummer	Datensatz	Beschreibung	in Version
1.1.15	f10eltern6	Die Variablen "f10e621e1"-"f10e621e3" wurden neu gelabeled.	1.1
1.1.16	f10eltern6	Geburtsmonate wurden verändert.	1.1
1.1.17	f10jugend	Bei den Fragen nach den Freizeitbeschäftigungen (f10j015c,e,f,g,i,o) wurden bei der Dateneingabe einzelne Items vertauscht, die aber einfach umgesetzt werden können. c=>n; d=>c; e=>d; f=>e; g=>f; h=>g; i=>h;n=>j; j=>i	1.1
1.1.18	f10lela	Durch einen Softwarefehler wurde die Frage nach der Anerkennung eines Abschluss im Ausland in der Screening Stichprobe nicht gestellt. Entsprechend ist die Variable "f10l038c" für diese Stichprobe auf "-2" zu setzen. Erst nach dem 18.6.2010 wurde diese Frage in der Kohorte gestellt - entsprechend werden auch hier alle anderen Werte auf "-2" gesetzt.	1.1
1.1.19	f10lela	Die Variable "sample" (jetzt "sample1") war mit falschen Werten belegt.	1.1
1.1.20	f10mihinc	Umkodierung der nicht-imputierten Werte auf -1 anstelle von -2 in den Originaldaten.	1.1
1.1.21	f10mipinc	Umkodierung der nicht-imputierten Werte auf -1 anstelle von -2 in den Originaldaten.	1.1
1.1.22	f10p	Die Kalenderdaten aus f10p070 waren in der ersten Datenlieferung nicht enthalten.	1.1
1.1.23	f10pgen	Die Variable "nation10" hat aufgrund falscher Bespaltung nicht die korrekten Ausprägungen.	1.1

Nummer	Datensatz	Beschreibung	in Version
1.2.1	alle	In allen Datensätzen bezeichnet "sample1" die Stichprobenzugehörigkeit. Ausnahmen sind ***f10pbrutto*** und ***f10kind*** ("psample") und ***f10hbrutto*** ("hsample")	**1.2**
1.2.2	alle	Sämtliche value Labels sind gleichlautend wie der Variablenname (einschließlich Groß- und Kleinschreibung).	1.2
1.2.3	alle	Die Ausprägungen in Variablen mit nur einem Wert wurden gelabled.	1.2
1.2.4	**Neue Gewichte**	Die Gewichte, die eine gemeinsame Hochrechnung von SOEP und FiD erlauben, befinden sich im Datensatz hhrf_fidsoep (Haushaltsebene) und phrf_fidsoep (Personenebene). Sie enthalten die Variablen zum Anspielen an andere Datensätze („hhnrakt“ sowie „persnr“) und jeweils drei Gewichtungsfaktoren pro Welle (in der Datenweitergabeversion 1.2 also nur drei). Diese drei Gewichtungsfaktoren ermöglichen eine Gewichtung einer kombinierten SOEP-FiD Population mit den Gewichtungsfaktoren „f10hhrf_soep“ (Haushalte) und „f10phrf_soep“ (Personen). Für den Fall, dass nur die FiD-Screening-Stichprobe bzw. nur die FiD-Kohorten-Stichprobe mit dem SOEP zusammen analysiert werden, sollten die jeweiligen Gewichte mit den Kürzeln „_scr“ (Screening) bzw. „_coh“ (Kohorte) verwendet werden.	1.2
1.2.5	bioage01	In den Variablen zu Störungen ("b01disord1"-"b01disord8") wurde noch der Code für "-1 no answer" hinzugefügt (vorher "-2").	1.2
1.2.6	bioage01	Die Variable "b01care" wurde in zwei Variablen "b01ccare" und "b01ccarei" umgesetzt.	1.2
1.2.7	bioage01	Die Variable "b01noccar10" wurde in "b01noccare10" umbenannt.	1.2
1.2.8	bioage01	Die Variable "b01breastf" wurde auf eine "1"/"2"-Kodierung umgestellt.	1.2
1.2.9	bioage01	Die Variable "b01breastfm" wurde für die Mütter, die momentan noch stillen auf "-2 Does not apply" gesetzt.	1.2
1.2.10	bioage01	Die Variable "b01nmedaid3m" wurde in "b01medaid3m" umbenannt.	1.2
1.2.11	bioage01	Die Variable "b01age" wurde neu kodiert: Nachdem Kinder vorher mindestens einen Monat alt waren, ist nun ein Alter von 0 Monaten möglich.	1.2

Nummer	Datensatz	Beschreibung	in Version
1.2.12	bioage01	Die Variable "b01suppartn" wurde ans SOEP angepasst und das Labeling geändert: "0" is nun "No partner in hh", "1" - "No support at all", "2" - "Some support", "3" - "Strong support", "4 - "Very strong support".	1.2
1.2.13	bioage01	Durch Fehler in der Kodierung wurden einige Werte in "b01preend" und "b01prebeg" geändert.	1.2
1.2.14	bioage02	Die Variable "b02breastf" wurde auf eine "1"/"2"-Kodierung umgestellt.	1.2
1.2.15	bioage02	Die Variable "b02breastfm" wurde für die Mütter, die momentan noch stillen auf "-2 Does not apply" gesetzt.	1.2
1.2.16	bioage02	Die Variable "b02age" wurde analog zu "b01age" neu kodiert.	1.2
1.2.17	bioage02	Die Variable "b02care" wurde in zwei Variablen "b02ccare" und "b02ccarei" umgesetzt.	1.2
1.2.18	bioage02	Für die Variable "b02breastfp" wurden missing codes ("-1", "-2") eingefügt.	1.2
1.2.19	bioage02	Für die Variable "b02lstmedex" wurde die Ausprägung "0" mit missing codes ("-1", "-2") codiert.	1.2
1.2.20	bioage02	Die Variablen "b02nactcar1", "b02nactcar2", "b02nactcar3" wurden in "b02nactcare1", "b02nactcare2", "b02nactcare3" umbenannt.	1.2
1.2.21	bioage03	Die Variablen "b03change1"-"b03change8" wurden für die Screening-Stichprobe auf "-2 Does not apply" gesetzt (vorher "0").	1.2
1.2.22	bioage03	Die Variable "b03care" wurde in zwei Variablen "b03ccare" und "b03ccarei" umgesetzt.	1.2
1.2.23	bioage03	Die Variable "b03breastf" wurde auf eine "1"/"2"-Kodierung umgestellt.	1.2
1.2.24	bioage03	Für die Variable "b03nmedaid" wurde jetzt eine "0" kodiert, wenn keine Arztbesuche vorlagen.	1.2
1.2.25	bioage03	Für die Variablen "b03hospital" und "b03hospital3m" wurde jetzt eine "0" kodiert, wenn kein Krankenhausaufenthalt vorlag. (In der Screening-Stichprobe bleibt die Variable "b03hospital3m" auf "-2 Does not apply".)	1.2
1.2.26	bioage03	Für die Variable "b03lstmedex" wurde die Ausprägung "0" mit missing codes ("-1", "-2") codiert.	1.2
1.2.27	bioage03	Die Variable "b03age" wurde analog zu "b01age" neu kodiert.	1.2

Nummer	Datensatz	Beschreibung	in Version
1.2.28	bioage06	Die Variablen "b06weight" und "b06height" waren durch einen Fehler beim Einlesen des *f10eltern4* Datensatzes falsch kodiert (siehe auch dort).	1.2
1.2.29	bioage06	Die Variablen "b06char6o", "b06char7o", "b06char8o" "b06char9o" wurden in "b06char6", "b06char7", "b06char8" und " b06char9" umbenannt.	1.2
1.2.30	bioage08p1/p2	Die Variable "b08waychi" wurde in "b08chhealth" umbenannt.	1.2
1.2.31	bioage08p1/p2	Die Variable "b08allownce" wurde gerundet auf ganze Zahlen.	1.2
1.2.32	bioage08p1/p2	In den Variablen "b08lamark" und "b08matmark" wurden die System-Missings auf "-1" gesetzt.	1.2
1.2.33	bioage08p1/p2	Durch eine neue Geschlechtszuordnung verschiebt sich ein Fall aus p1 in p2 und umgekehrt, wodurch sich in einem Fall jeweils mehrere Variablenausprägungen ändern.	1.2
1.2.34	bioage10p1/p2	Die Variablen "b10freqfre1"-"b10freqfre14" wurden in "b10freqact1"-"b10freqact14" umbenannt.	1.2
1.2.35	bioage10p1/p2	Die Variable "b10allownce" wurde gerundet auf ganze Zahlen.	1.2
1.2.36	bioage10p1/p2	In den Variablen "b10lamark" und "b10matmark" wurden die System-Missings auf "-1" gesetzt.	1.2
1.2.37	bioage10p1/p2	Durch eine neue Geschlechtszuordnung verschiebt sich ein Fall aus p1 in p2 und umgekehrt, wodurch sich in einem Fall jeweils mehrere Variablenausprägungen ändern.	1.2
1.2.38	biobirth	Der Datensatz wurde auf Personen mit einem Personeninterview sowie die Jugendlichen beschränkt, nachdem vorher versucht wurde, aus anderen Informationen im Haushalt auf eigene Kinder rückzuschließen.	1.2
1.2.39	biocouply	Durch Verbesserungen in der Kodierung der Partnerschaftsbeziehungen (Berücksichtigung der Nicht-Antwortenden) erhöhen sich die Fallzahlen.	1.2
1.2.40	biomarsy	Durch Verbesserungen in der Kodierung der Partnerschaftsbeziehungen (Berücksichtigung der Nicht-Antwortenden) erhöhen sich die Fallzahlen.	1.2

Nummer	Datensatz	Beschreibung	in Version
1.2.41	f10hbrutto	Einige Fälle (19) waren in "f10bula" als "Rheinland-Pfalz" codiert, kommen aber eigentlich aus dem "Saarland".	1.2
1.2.42	f10hbrutto	Die Variable "f10intza" (Zahl der eingesetzten Interviewer) wurde eingeführt.	1.2
1.2.43	f10hgen	Die Variable "cnstry10" wurde umbenannt in "cnstyr10".	1.2
1.2.44	f10hgen	Umbenennung der Variablen "elec10" in "electr10", sowie "felec" in "felectr10" (SOEP Standard seit 2010).	1.2
1.2.45	f10hgen	Die imputierten Variablen ("size10", "room10", "rent10", "heat10", "util10", "electr10", "i1hinc10", "i2hinc10", "i3hinc10", "i4hinc10", "i5hinc10") haben durch die neuen Imputationen neue Werte. Außerdem wurde die Impuationsroutine für "util10" geändert, so dass "futil10" zusätzliche Impuationen aufweist.	1.2
1.2.46	f10hgen	Einige Fälle (89) waren in "nuts10" als "Rheinland-Pfalz" codiert, kommen aber eigentlich aus dem "Saarland" (siehe ***f10hbrutto***).	1.2
1.2.47	f10hgen	Zur Variable "typ2hh10" wurde die Ausprägung "73 Grandparent(s)-grandchild(ren) hh" hinzugefügt, so dass die Abgrenzungen genauer werden. Dadurch verändert sich auch typ1hh10 in einigen Fällen.	1.2
1.2.48	f10hgen	Bei der Berechnung von "ahinc" wurde jetzt auch das Einkommen aus Vermietung und Verpachtung berücksichtigt (betrifft 130 Fälle).	1.2
1.2.49	f10hgen	Die Variablen "singpa", "lrgfam" und "lowinc" wurden als stichprobenspezifische und wellenunveränderliche Variablen in ***hpfad*** überführt.	1.2
1.2.50	f10intview	In der Variable "f10dura" wurden alle Elternfragebögen von system-missing (".") auf "-2" gesetzt.	1.2
1.2.51	f10jugend	Die Ausprägungen der Variablen "f10j062", "f10j066b", "f10j068", "f10j079d" wurden gelabeled.	1.2
1.2.52	f10kind	Es wurden einige Geburtsmonatsveränderungen mit aufgenommen (Vergleich mit ***f10pbrutto***).	1.2
1.2.53	f10kind	Die Indikatoren der Mutter- und Vaterzeiger wurden verbessert, und benutzen nun auch die Informationen aus den Beziehungenhistorien (***biocouply***). Dadurch konnten zahlreiche unbekannte Eltern-Kind-Beziehungen zugeordnet werden.	1.2

Nummer	Datensatz	Beschreibung	in Version
1.2.54	f10lela	Die Ausprägungen Variablen "f10l003" (Geburtsland), "f10l008b" (Land zweiter Staatsangehörigkeit) und "f10l010" (Alte Staatsangehörigkeit) waren nicht gelabled.	1.2
1.2.55	f10lela	Umkodierung von Beziehungsvariablen im Haushalt mit der "hhnrakt" 2000876 (siehe *f10pbrutto*).	1.2
1.2.56	f10lela	Umkodierung der Variable "f10l064e2" auf 2003 im Haushalt mit der "hhnrakt" 2005525.	1.2
1.2.57	f10mihinc	Verbesserte Imputationen.	1.2
1.2.58	f10mipinc	Verbesserte Imputationen, insbesondere Imputationen der nicht-antwortenden Haushaltsmitglieder (PUNRs). Diese sind durch die zusätzliche Variable "punr" gekennzeichnet.	1.2
1.2.59	f10mipinc	Die Variablen "f10pchdsup" bzw. "i_f10pchdsup" wurden in "f10palimon" bzw. "i_f10palimon" umbenannt.	1.2
1.2.60	f10p	Der Variable "f10p102n" (Sonstige Sorgen) wurden labels hinzugefügt.	1.2
1.2.61	f10pbrutto	Die Variable "f10stell" wurde im Haushalt mit der "hhnrakt" 2000876 geändert.	1.2
1.2.62	f10pbrutto	Die Variable "f10hhnrold" wurde für alle Fälle von "0" auf "-2" gesetzt.	1.2
1.2.63	f10pgen	Die Variablen "agrhrs10", "acthrs10" und "ovrhrs10" wurden mit einem neuen Umrechnungsfaktor für Wochenstunden berechnet.	1.2
1.2.64	f10pgen	Durch Verbesserungen in biomars und biocouply wurden die Variablen "partp10", "partno10", "coupst10", "coupid10" und "marrst10" in einigen Fällen verändert.	1.2
1.2.65	f10pgen	Durch genauere Kodierung der Abschlüsse in der ehemaligen DDR haben sich in den Variablen "scedu10", "scedue10", "vcdege10", "colleg10", "timedu10", "casmin10" und "isced10" einige Fälle verändert.	1.2
1.2.66	f10pgen	Durch falsche Kodierung in der Variable "occpos10" waren 486 Individuen falsch als Rentner (Wert "13") kodiert, die entweder nicht erwerbstätig ("10") sind, keine Angabe gemacht haben ("-1") oder auf die die Frage nicht zutrifft ("-2"). Entsprechend ändert sich auch "egp10".	1.2
1.2.67	f10pgen	Die arbeitenden Jugendlichen bekommen jetzt einen Erwerbstatus zugewiesen, so dass sich die Variable "lfs10" ändert.	1.2

Nummer	Datensatz	Beschreibung	in Version
1.2.68	f10pgen	Durch die neuen Imputationen verändern sich die Variablen "labgro10" und "labnet10".	1.2
1.2.69	hbrutt10_fid	Die Variablen "f10apro", "f10wuma1", "f10wuma2", "f10wuma3", "f10wuma401", "f10wuma402", "f10wuma403", "f10wuma404", "f10wuma4ka", "f10wuma5", "f10wuma6" und "f10wuma7" wurden für die Screening Stichprobe auf "-2 Trifft nicht zu" gesetzt (vorher "0").	1.2
1.2.70	hpfad	Die Variablen "singpa" (Alleinerziehend bei erstem Kontakt), "lrgfam" (Mehrkindfamilie bei erstem Kontakt) und "lowinc" (Niedrigeinkommenshaushalt bei erstem Kontakt) wurden aus ***f10hgen*** in ***hpfad*** übernommen, weil es sich hierbei um Informationen handelt, die sich nicht über die Wellen ändern und außerdem Informationen über die Stichproben geben.	1.2
1.2.71	ppfad	Durch Verbesserungen in den Mutter- und Vaterzeigern konnten zahlreiche unbekannte Beziehungen in den Variablen "germborn", "immiyear", "corigin" und "migback" neu zugordnet werden.	1.2
1.2.72	ppfad	In Vorbereitung auf die neue Welle bekommt die Variable "f10netto" neue labels: "110 FiD: Interviewee P-Interview", "111 FiD: Interviewee P-Interview & Biography I", "112 FiD: Interviewee P-Interview & Biography II", "113 FiD: Interviewee P-Interview & Biography I+II"	1.2

Nummer	Datensatz	Beschreibung	in Version
2.0.1	bioage01-10	Die Sprachen, in denen mit den Kindern im Haushalt gesprochen wird, sind nun codiert nach ISO-639-1 Standard in den Datensätzen enthalten als "$language01"-"$language03".	**2.0**
2.0.2	bioage01-10	Die Variablen "care6", "care8", und "care8h" wurden wegen Problemen in der Befragung neu berechnet. Eine genauere Beschreibung findet sich in der bioage01-10 Dokumentation.	2.0
2.0.3	bioage01-10	Die Variablen "care1h"-"care12h" wurden so berechnet, dass die Gesamtsumme der Betreuungszeit maximal 168 Stunden beträgt.	2.0
2.0.4	bioage01-10	Es wurden einige Variablennamen in den ***bioage*** files geändert; sämtliche neuen Namen finden sich in der entsprechenden Tabelle in der bioage01-10 Dokumentation. Ebenso wurden einige Variablen und Value Labels geändert, um eine bessere Benutzung zu ermöglichen.	2.0
2.0.5	bioage06 - bioage10	In diesen drei ***bioage*** Datensätzen wurde die Variable "bXXage" (Alter des Kindes in Monaten) hinzugefügt, so dass jetzt sämtliche ***bioage*** files diese Variable enthalten.	2.0
2.0.6	bioage08, bioage10	In den Variablen "scoldura", "hospital" und "nmedaid" wurden kleinere Codierungsfehler beseitigt.	2.0
2.0.7	biomarsy	Durch neu eingeführte Fragen im P-Fragebogen nach gleichgeschlechtlichen Partnerschaften haben sich auch die "spelltyp" Variablen in biomarsy und biocouply entsprechend geändert.	2.0
2.0.8	f10eltern1	Die Benennung der Variablen aus der Frage 28 (Betreuungseinrichtungen/-personen) wurde an das FiD-Prinzip angeglichen. Vorherige Variablen "f10e128a"-"f10e128a28o" heißen jetzt "f10e128a1", "f10e128a2" bis "f10e128g1", "f10e128g2", "f10e128h".	2.0
2.0.9	f10eltern1-f10eltern3	Für die U-Untersuchungen in "f10e109a/b", "f10e212a/b" und "f10e312a/b" wurden Value Labels eingesetzt.	2.0
2.0.10	f10h	Die Variablen "f10h042a"-"f10h042u" (Transferleistungen im letzten Jahr) wurden in ihrer Benennung an die FiD-Logik angepasst, und in "f10h042a1", "f10h042a2" bis "f10h042g3" umbenannt. Gleiches gilt für "f10h044a"-"f10h044o" (aktuelle Transferleistungen), die zu "f10h042a1"-"f10h042g2" umbenannt wurden.	2.0

Nummer	Datensatz	Beschreibung	in Version
2.0.11	f10hbrutto	Die Variable f10mkz2 (Migrantenkennzeichen) wurde neu kodiert, "0" wird zu "1", "1" zu "2", "2"zu "3".	2.0
2.0.12	f10hgen	Die Variable "rent10" war durch einen Kodierungsfehler falsch berechnet. Sie ist jetzt korrigiert und sollte ausschließlich benutzt werden.	2.0
2.0.13	f10hgen	Durch neue Imputationen haben sich alle imputierten Werte ("rent10", "heat10", "util10", electr10, i_hinc10, size10, room10) verändert.	2.0
2.0.14	f10lela	Das Label des Wertes "1 Trennung" in der Variable "f10l062f" (Ende der Beziehung) von auf "2 Scheidung" geändert.	2.0
2.0.15	f10lela	In der Variable "f10l062f" (Ende der Beziehung) wurden drei Fälle nachträglich auf "-2 Trifft nicht zu" umgesetzt, deren Ehe noch Bestand hatte ("f10l062d"=1), die aber aus technischen Gründen Scheidung bzw. Tod des Ehepartners angegeben hatten.	2.0
2.0.16	f10mihinc	Mit den neu berechneten Imputationen ergeben sich auch neue Werte für "i_hinc" in ***hgen***, so dass ausschließlich ***f10hgen*** der neuen Distribution genutzt werden sollte.	2.0
2.0.17	f10mipinc	Die Variable "f10plsyrinc" (imputiertes Einkommen des letzten Jahres) wurde bei den neuen Imputation nicht mehr berücksichtigt.	2.0
2.0.18	f10mipinc	Mit den neu berechneten Imputationen ergeben sich auch neue Werte für "labgro10" und "labnet10" in ***f10pgen***, so dass ausschließlich ***f10pgen*** der neuen Distribution genutzt werden sollte.	2.0
2.0.19	f10p	In der Variable "f10p030" (Art des Arbeitsvertrags) wurden die Value Labels korrigiert, auf jetzt "1 unbefristet" und "2 befristet".	2.0
2.0.20	f10pgen	Wegen identisch benannter Variablen wurde "fsize10" in "cosize10" umbenannt, sowie "fcsize10" in "crsize10".	2.0
2.0.21	f10pgen	Die Variablen "expft10", "exppt10" und "expue10" sowie "tenure10" sind rückwirkend aus den Daten aus 2011 berechnet und eingeführt worden.	2.0
2.0.22	f10pgen	Durch neue Imputationen haben sich die imputierten Einkommen ("labgro10", "labnet10") verändert.	2.0

Nummer	Datensatz	Beschreibung	in Version
2.0.23	f10pgen	Durch neu eingeführte Fragen im P-Fragebogen nach gleichgeschlechtlichen Partnerschaften haben sich auch die Variablen "marrst10" und "coupst10" entsprechend geändert.	2.0
2.0.24	f10pkal	Die Variable "f10p1o" (Short-Time working Jan-Dec 2009) und die dazugehörigen Variablen "f10p1o001"-"f10p1o012" wurden umbenannt in "f10p1k" bzw. "f10p1k001"-"f10p1k012". Damit heißen diese Variablen wie im SOEP 2010.	2.0
2.0.25	f10pkal	Die Indikator-Variablen " f10p1a01"-"f10p1n01" sowie die dazugehörigen Monatsangaben "f10p1a02"-" f10p1n02" wurden in die Weitergabe mit aufgenommen.	2.0
2.0.26	hhrf/phfr	Durch die Hinzunahme der Screening Stichprobe 2011 wurde eine Unterscheidung der Hochrechnungsfaktoren für die Screening Stichproben nötig. Daher gibt es jetzt die Variablen "f10phrf_sc10" (Screening 2010), "f10phrf_ch10" (Kohorte) und "f11phrf_sc11" (Screening 2011). Entsprechend auch für die Variablen "hhrf_sc10", "hhrf_ch10" und "hhrf_sc11".	2.0

Nummer	Datensatz	Beschreibung	in Version
2.1.1	**Neue Gewichte**	In dieser Version der Datenweitergabe gibt es erneut Gewichte, die eine gemeinsame Nutzung von FiD und SOEP ermöglichen. Sie finden sich in den Datensätzen hhrf_fidsoep und phrf_fidsoep. Es werden darüber hinaus Querschnittsgewichte für 2011 für die FiD-Stichproben bereitgestellt, die eine Analyse der FiD-Daten alleine ermöglichen. Diese Hochrechnung erfolgte nach einem komplett neuen Verfahren. Die Dokumentation der Gewichte enthält nähere Details.	**2.1**
2.1.2	bioage01-03	Die Variablen "health1"-"health8" wurden umbenannt: "health1" heißt jetzt "health", und ist bezieht sich auf die Sorgen der Mutter um die Gesundheit des Kindes. "health2"-"health6" wurden zu "temp1"-"temp5", "health7" wurde zu "temp7" und "health8" wurde zu "temp6". Damit folgt FiD der Logik im SOEP Datensatz ***bioagel*** (siehe 2.1.5), der Längsschnittversion der Bioagedaten im SOEP.	2.1
2.1.3	bioage10	Die Variablen "imper1a", "imper2a" und "imper3a" hatten bisher Werte von 3, die eigentlich als "-2 Does not apply" hätten kodiert werden müssen. Dies wurde jetzt korrigiert.	2.1
2.1.4	bioage10	In sämtlichen Variablen "imperXa"-"imperXe" waren bisher Werte auf "-2 Does not apply", auch wenn es sich um eine sogenannte "gesammelte fehlende Angabe" handelte, die Person also die Antwort verweigert hatte. Dies war in "imperXf" als "-1 No answer" festgehalten. Jetzt wurden bei diesen Fällen sämtliche Antworten auf "-1 No answer" gesetzt.	2.1
2.1.5	bioagel	Dieser Datensatz wurde neu in die Weitergabe aufgenommen. Er enthält die kombinierte Fassung der ***bioage01-bioage10*** Datensätze. Variablen werden hier ohne ihren prefix (bXX) angegeben, so dass Variablen, die über die Fragebögen hinweg identisch erhoben wurden, direkt für ein Kind verglichen werden können.	2.1

Nummer	Datensatz	Beschreibung	in Version
2.1.6	biobirth	Der Datensatz ***biobirth*** wurde komplett neu aufgesetzt. Dadurch kommt es zu Veränderungen, hauptsächlich gibt es zahlreiche neue Fälle, weil erstmals auch nicht-teilnehmende Personen berücksichtigt wurden. Es wurde außerdem die Variable "kidhome" entfernt, weil sie zeitveränderlich ist; ob ein Kind im Haushalt ist, lässt sich über seine Personennummer und den jahresspezifischen ***pbrutto*** Datensatz anspielen. Außerdem wird jetzt zur höheren Genauigkeit die Variable "kidsource" einzeln für jedes Kind aufgeführt.	2.1
2.1.7	biocouply/ biomarsy	Durch leichte Veränderungen in der Kodierung (die SOEP-Änderungen folgen) ist es zu einer Erhöhung der Fallzahlen um rund 1000 Spells in beiden Datensätzen gekommen. Dadurch haben sich auch manche Variablen geändert.	2.1
2.1.8	f10pgen/ f11pgen	Die Kodierung der variablen "mps$$" war fehlerhaft und wurde jetzt korrigiert.	2.1
2.1.9	f10pgen/ f11pgen	In der Weitergabeversion 2.0 waren die Variablenlabels der Variablen "partp$$" und "partno$$" falsch und wurden geändert.	2.1
2.1.10	f10pgen/ f11pgen	In der Variable "coupst10" (Partner Status) waren 4 Fälle mit gleichgeschlechtlicher Partnerschaft falsch zugeordnet, in "coupst11" betraf dies 5 Fälle.	2.1
2.1.11	f10pgen/ f11pgen	Durch die Veränderungen in ***biocouply*** (siehe 2.1.7) hat sich für gut 80 Fälle die Variable "marrst" von "3 single" auf "4 geschieden" verändert.	2.1
2.1.12	f10pgen/ f11pgen	Die Variable "tenure10"/"tenure11" (Betriebszugehörigkeit) wurde in wenigen Fällen geändert. Fälle mit unbekanntem Monat des Arbeitsbeginns hatten den Wert "-1" und bekommen nun einen Wert, sofern das Jahr des Arbeitsbeginns bekannt ist. Der entsprechende Monat wurde dabei zufällig besetzt.	2.1
2.1.13	f10pkal/ f11pkal	Bei den Variablen "f10p2n01" und "f11p2n01" (keine dieser Einkunftsarten im Vorjahr) wurden die Werte nach SOEP-Logik umgesetzt: "-2 Trifft nicht zu" wird gesetzt wenn eine Einkommensart vorliegt.	2.1
2.1.14	f10pkal/ f11pkal	In den Variablen "p1a01"-"p1n01" waren die Werte "1 Yes" und "2 No" vertauscht, was in der neuen Version behoben wurde.	2.1

Nummer	Datensatz	Beschreibung	in Version
2.1.15	f11h	Die Variable "f10h034b" (Absetzen von Verlusten im Vorjahr) war mit "f10" falsch benannt und wurde jetzt korrigiert auf "f11h034b".	2.1
2.1.16	f11hbrutto	Die Variable "f11hhnrold" (HH-Nummer im Vorjahr) wurde dem Datensatz jetzt hinzugefügt. Bisher war diese Information ausschließlich im Datensatz ***hpfad*** enthalten.	2.1
2.1.17	f11pbrutto	Die Variable "f11pnrold" (laufende Personennummer im Vorjahr) wurde jetzt auf maximal 2 Stellen reduziert und ist so mit der aktuellen laufenden Nummer vergleichbar. (Dies betrifft in keiner Weise die unveränderliche Personennummer, "persnr".)	2.1
2.1.18	hhrf / phfr / hhrf_fidsoep / phfr_fidsoep	Zur Verbesserung der Übersichtlichkeit wird ab Version 2.1 darauf verzichtet, spezifische Hochrechnungsfaktoren für die Kohorten- bzw. Screening-Stichprobe bereitzustellen. Bei Bedarf können sie auf individuelle Nachfrage zur Verfügung gestellt werden.	2.1
2.1.19	ppfad	Durch die Veränderungen in ***biobirth*** (siehe 2.1.6), denen verbesserte Zuordnungen von Kindern zu ihren Eltern zugrunde liegen, kommt es bei den Variablen zum Migrationshintergrund ("germborn", "immiyear", "corigin", "migback" und "miginfo") zu Unterschieden zur Vorversion. Diese beziehen sich ausschließlich auf Kinder und nicht-teilnehmende Personen.	2.1
2.1.20	f10paradata / f11paradata	In den Datensätzen zur Interviewsituation wurde eine neue Variable hinzugefügt, die den Wochentag der Befragung angibt (f$dow).	2.1
2.1.21	f10p/f11p	Die kodierten Berufe und Branchen, die bisher in ***f10pgen*** und ***f11pgen*** vorlagen, wurden nun auch in ***f10p*** und ***f11p*** hinzugefügt (Variablen "f10p021_is88", "f10p021_klas", "f10p060_is88n", "f10p060_klasn", "f10p026_nace", "f11p022_is88", "f11p022_klas", "f11p060_is88n", "f11p060_klasn", "f11p027_nace").	2.1

Nummer	Datensatz	Beschreibung	in Version
2.1.22	f10p/f11p	Die Variablen zum Einkommen im Vorjahr ("f10p071" und "f11p075") wurden an die SOEP Benennung angepasst. Dabei wurden die Buchstaben, die das Einkommen kennzeichnen, geändert. Unter anderem erhalten die FiD-spezifischen Variablen zu Kindesunterhalt (jetzt "u" statt "i"), Betreuungsunterhalt ("v" statt "j") und nachehelicher Unterhalt ("w" statt "k") neue Zuordnungen.	2.1
2.1.23	f10p/f11p	Die Variablen zu Sondervergütungen ("f10p072" und "f11p076" wurden jetzt nach der FiD-Logik benannt, d.h. mit Buchstaben für das Item sowie Zahl für Antworten innerhalb des Items.	2.1
2.1.24	mehrere Datensätze	Die Variable, die angibt, durch welchen Interviewer das Interview durchgeführt wurde, wurde in allen Datensätzen, sofern die Information vorkommt, einheitlich in "intid" umbenannt. Diese Änderung entspricht der neuen Benennung im SOEP, Version 28.	2.1
2.1.25	f10kind/f11kind	Die Veränderungen in ***biobirth*** (siehe 2.1.6) führen zu einem Update der Elternzeiger, also der Variablen "mothno$$", "mothp$$", "fathno$$" sowie "fathp$$".	2.1
2.1.26	alle Datensätze	Es wurden - wie im SOEP, Version v28 - neue Missing-Codes eingeführt. Sie enthalten zusätzliche Informationen, wenn Interviewten einzelne Fragen nicht gestellt wurden. Im Einzelnen handelt es sich um folgende neue Ausprägungen: -4: "Invalide Mehrfachnennung" -5: "Frage in Sample nicht gestellt" -6: "Sample-spez. Filter"	2.1
2.1.27	f10jugend / f11jugend	Die Variablen "f$j037g", "f$j037h", "f$j037i" wurden hinzugefügt. Sie geben das Leistungsniveau von Gesamtschülern in den Hauptfächern an.	2.1

Nummer	Datensatz	Beschreibung	in Version
3.0.1	Alle Elterndatensätze aller Wellen	Die Variable zum Interview-Mode wurde in 2011 von "f11pinta" in "f11e#inta" umbenannt. 2010 wurde sie hinzugefügt ("f10e#inta").	**3.0**
3.0.2	f12eltern6	Die Fragen zum Kind (Stärken und Schwächen, SDQ) wurden im Fragenbogen Eltern 6 randomisiert mit verschiedenen Skalen (7-Punkt und 3-Punkt) abgefragt (siehe auch 3.0.4). Darüber hinaus haben "f12e6a23a"-"f12e6a23y" (3-Punkt-Skala) sowie "f12e6b23a"-"f12e6b23y" (7-Punkt-Skala) ab 2012 den vollen Umfang von 25 Variablen. In 2010 und 2011 wurden im Elternfragebogen 6 nur 18 Variablen ("f11e622a"-"f11e622r") abgefragt. Die neuen Variablen in ***f12e6*** sind benannt als "f12e6#23s"-"f12e6#23y", was nicht der Reihenfolge im Fragebogen entspricht.	3.0
3.0.3	f12eltern6	In ***f12e6*** ist die Frage zu den schulischen Aktivitäten des Kindes außerhalb des Unterrichts ("f12e614a"-"f12e614d") neu. Daher wurden die folgenden Fragen nach hinten verschoben. Zudem wurden die Fragen zu den Freizeitbeschäftigungen des Kindes ("f12e615a"-"f12e615n") und zum besonderen pädagogischen Förderbedarf ("f12e616a"-"f12e616i") vertauscht. Die Fragen "f12e615a"-"f12e615n" entsprechen demnach "f11e618a"-"f11e618n" und die Fragen "f12e616a"-"f12e616i" entsprechen "f11e614a"-"f11e614i".	3.0
3.0.4	bioage10	Die Fragen zum Kind (Stärken und Schwächen, SDQ) wurden im Fragenbogen Eltern 6 mit verschiedenen Skalen (7-Punkt und 3-Punkt) abgefragt, um die international übliche 3-Punkt Skala auch für die vorherigen Jahre anwenden zu können. Entsprechend haben sich für alle Jahre die BEHAV Variablen auf die 3-Punkt-Skala reduziert. Nähere Information zum Verfahren finden sich in der Dokumentation zu ***bioage10***. Diese Änderung findet sich auch im Längsschnittdatensatz ***bioagel***.	3.0

Nummer	Datensatz	Beschreibung	in Version
3.0.5	f12p	Es gibt im Datensatz ***f12p*** (Personenfragebogen 2012) einige Personen, die vorher noch den gesamten Biographiefragebogen (oder einen Teile) beantworten. Folgende Variablen liegen für diese Personen im ***f12lela*** vor, während sie im ***f12p*** den Wert "-6 Sample-spez. Filter" annehmen: ***f12lela*** ***f12p*** "f12l002" "f12p122" "f12l006" "f12p125" "f12l007" "f12p129" "f12l008" "f12p126" "f12l009" "f12p127" "f12l045" "f12p137" "f12l046" "f12p138" "f12l061" "f12p116" "f12l063" "f12p117" "f12l064" "f12p118" "f12l065" "f12p119" "f12l066" "f12p120"	3.0
3.0.6	f11luecke	In dieser Welle wird erstmals der Datensatz ***f11luecke*** weitergegeben, der für diejenigen Personen, die in einem Jahr nicht teilgenommen haben, eine Kurzinformation über dieses Jahr erhoben wird. Hier werden Informationen über das Jahr 2011 erhoben, und die Daten daher im Ordner „2011“ abgelegt.	3.0
3.0.7	f12pbrutto / f12kind	Analog zum SOEP wurde in FiD die variable "$stell" (bzw. "$kstell" in ***$kind***) in 2012 umgestellt, um die Beziehungen der Personen im Haushalt zum Haushaltsvorstand besser erfassen zu können als bisher. Sämtliche Codes der Vorwellen sind direkt abbildbar, eine Tabelle zur Überführung findet sich in der Dokumentation zu ***$kind***.	3.0
3.0.8	$pgen	Die Werte zur Berufserfahrung ("expft$$", "exppt$$", "expue$$") wurden neu codiert, was zu Änderungen in den 2010 gezogenen Stichproben führte. Außerdem liegen nun erstmals Informationen für die 2011 gezogenen Screening Stichprobe vor.	3.0
3.0.9	$pgen	Die Berechnung der Variable "tenure$$" (Betriebszugehörigkeit) wurde neu codiert, was zu leichten Veränderungen führt.	3.0

Nummer	Datensatz	Beschreibung	in Version
3.0.10	$pgen	Die Variablen rund um die Berufsklassifizierung ("class$$", "is88$$", "nace$$", "isei$$", "egp$$", "siops$$", "mps$$") werden ab dieser Welle mit den Werten der Vorwelle ersetzt, wenn diese vorliegen. Aufgrund der Filterführung werden diese Variablen nur alle zwei Jahre erfragt, sofern kein Berufswechsel angegeben wurde.	3.0
3.0.11	$pgen	Fehlende Werte für die Variablen "scedu11", "scedue11", "scedua11", "vcdeg11", "vcnone11", "colleg11", "timedu11", "isced11", "casmin11" aus 2011 wurden durch die nachträgliche Erhebung des ersten Biographieteils in 2012 für die Personen umgesetzt, die auch in 2012 noch an der Befragung teilnehmen. (Personen, die 2011 erstmals in einem 2010 gezogenem Haushalt befragt wurden, wurde fälschlicherweise der 2. Biographieteil gegeben.)	3.0
3.0.12	f12pkal	Die Variable "f12p2m03" ist ausschließlich mit missings besetzt und wird auf "-2" gesetzt. Die entsprechende Variable im P-Fragebogen "f12p084o3" wurde in 2012 nicht erhoben. Möglichkeit zur Nachbesserung besteht über Frage "f12p088".	3.0
3.0.13	f10mihinc / f11mihinc	Neu berechnete Imputationen, die nun auch die in 2012 erhobenen Daten berücksichtigen. Mit den neu berechneten Imputationen ergeben sich auch neue Werte für "i_hinc" in ***$hgen***, so dass ausschließlich die neue Distribution genutzt werden sollte.	3.0
3.0.14	f10mipinc / f11mipinc	Neu berechnete Imputationen, die nun auch die in 2012 erhobenen Daten berücksichtigen. Mit den neu berechneten Imputationen ergeben sich auch neue Werte für "labgro10" und "labnet10" in ***$pgen***, so dass ausschließlich die neue Distribution genutzt werden sollte.	3.0
3.0.15	f12paradata	Mit der Aufnahme der ***$luecke*** Datensätze (siehe auch 3.0.6) gibt es einen zusätzlichen Code in der Variable "qstnr", und zwar "10 Luecke-Qunaire (Gap in prev. year)". Obwohl der Datensatz ***$luecke*** inhaltlich zur Vorwelle gehört, wird er in ***$paradata*** in der Erhebungswelle (hier f12) abgelegt.	3.0

Nummer	Datensatz	Beschreibung	in Version
3.1.1	**Neue Gewichte**	In dieser Version der Datenweitergabe gibt es erneut Gewichte, die eine gemeinsame Nutzung von FiD und SOEP (Jahre 2010-2012) ermöglichen. Diese Hochrechnungsfaktoren sind in den Datensätzen ***hhrf_fidsoep*** und ***phrf_fidsoep***. Es werden darüber hinaus Querschnittsgewichte für 2011 und 2012 für die FiD-Stichproben bereitgestellt, die eine Analyse der FiD-Daten alleine ermöglichen. Außerdem gibt es eine rückwirkende Änderung aller Gewichte aufgrund einer Umstellung der Hochrechnungsfaktoren der Kohorten-Haushalte mit Kindern, die 2010 geboren wurden. Diese Umstellung wird im Dokument **"NeueGewichte2012.pdf"** erläutert. Weitere Details sind in der Dokumentation der Gewichte enthalten.	**3.1**
3.1.2	bioage files	Im Zuge der Angleichung an das SOEP wurden mehrere Variablen in allen ***bioage*** Files umbenannt: PERSNRM -> PERSNRRESP (persnr der ausfüllenden Person) SVYYEAR > SYEAR (Jahr der Befragung) NMEDAID -> MEDAID3M (Ärztliche Hilfe in den letzten 3 Monaten) HOSPITAL -> HOSPITAL12M (Krankenhausaufenthalte in den letzten 12 Monaten) Andere Änderungen beziehen sich auf einzelne Datensätze (siehe 3.1.3-3.1.7).	3.1
3.1.3	bioage01	In den Variablen "B01PREGMO" (Schwangerschaftsmonat der Mutter bei Interview im Vorjahr) sowie PREGMY kam es durch einen Fehler in der Kodierung zu falschen Werten in 83 Fällen aus 2012. Dieser Fehler wurde nun behoben.	3.1
3.1.4	bioage01, bioage03, bioage03	Im Zuge der Angleichung an das SOEP wurde die Variable MEDAID3M (ärztliche Hilfe in den ersten 3 Monaten nach der Geburt) in MEDAID3MB umbenannt. Nutzer sollten darauf achten, diese Variable nicht mit MEDAID3M (ehemals NMEDAID) zu verwechseln.	3.1

Nummer	Datensatz	Beschreibung	in Version
3.1.5	bioage08, bioage10	Bisher wurde in diesen beiden bioage Files die Variable ALLOWNCE (Taschengeld im Monat) nur auf Monatsebene weitergegeben. In Übereinstimmung mit dem SOEP wird nun auch das wöchentliche Taschengeld berechnet. Nun gibt es zwei Variablen: ALLOWPM (Taschengeld pro Monat) und ALLOWPW (Taschengeld pro Woche).	3.1
3.1.6	bioage08, bioage10	In Übereinstimmung mit dem SOEP wurden die Variablen CURSCOL5 (noch nicht in der Schule) durch SCLENROLN, STSCOLYR durch SCLENROLY, und STSCOLMN durch SCLENROLM ersetzt. Außerdem wurden die Variable CURSCOL4 (sonstige Schule) in CURSCOL8 umbenannt, und die Variablen CURSCOL4 (Hauptschule), CURSCOL5 (Realschule), CURSCOL6 (Gymnasium) und CURSCOL7 (Gesamtschule) hinzugefügt, wobei letztere (anders als beim SOEP) aus dem Haushaltsfragebogen gewonnen werden.	3.1
3.1.7	bioage10	In Übereinstimmung mit dem SOEP wurde die Variable CONSCHO5 (Kein Kontakt mit der Schule) in CONSCHO7 umbenannt. Außerdem wurden die Variablen ALLOW (Taschengeld) und NOMARK (keine Noten) hinzugefügt.	3.1
3.1.8	biojob	Aufgrund von Unklarheiten bei der Generierung des Files biojob wurde er aus der Weitergabe 3.1 herausgenommen. Auf Anfrage kann er Nutzern zur Verfügung gestellt werden.	3.1
3.1.9	f10jugend	Die Werte der Variablen F10J015I, F10J015J und f10J015O waren untereinander vertauscht und wurden nun korrigiert.	3.1
3.1.10	f12jugend	In den Variablen F12J037G und F12J037H waren in der Version 3.0 falsche Labels vergeben worden ("[1] -1"; "[2] A"; "[3] B"; "[4] C"). Dies wurde nun korrigiert, so dass nun "[-1] -1", "[1] A", "[2] B", und "[3] C" deklariert wurden.	3.1
3.1.11	f11hgen	Bei der Berechnung der Variablen AHINC11 wurde bei 36 Fällen statt des Witwenpensionen eine falsche Komponente addiert. Die Unterschiede sind für diese Fälle gering.	3.1

Nummer	Datensatz	Beschreibung	in Version
3.1.12	f12hgen	Bei der Berechnung der Variablen AHINC12 kann im Gegensatz zu den Vorjahren nicht mehr auf die Pensionen zurückgegriffen werden, weil diese Frage nicht mehr gestellt wurde. Entsprechend erfolgt die Berechnung von AHINC12 nun ohne diese Komponente. Irrtümlicherweise war in der Version 3.0 der Fehler aus 2011 (siehe 3.1.11) wiederholt worden, so dass nun 34 Fälle korrigiert wurden.	3.1
3.1.13	f12eltern2	Für die Person mit der PERSNR 21504004 war in der Weitergabe 3.0 in der Variable F12E205G irrtümlich der Wert "6" angegeben. Dieser wurde nun auf "-1 Keine Angabe" umgesetzt.	3.1
3.1.14	$kind	Durch Verbesserungen bei der Partnerdefinition kann es vereinzelt zu Umcodierungen bei den Variablen MOTHP$$ und FATHP$$ (Indikator für das Verhältnis zwischen Kind und Eltern) kommen, die die Information "6 - unbekannt" revidieren.	3.1
3.1.15	f12kind	Folgende Variablen wurden in der Weitergabe 3.0 aufgrund von Verschiebungen im Fragebogen falsch benannt und nun korrigiert: F12K056 -> F12K061 F12K057a, b -> F12K062a, b F12K058 -> F12K063 F12K059 -> F12K064 F12K060 -> F12K065 F12K061a, b -> F12K066a, b F12K062 -> F12K067 F12K063 -> F12K068 F12K064a, b -> F12K069a, b F12K065a, b -> F12K070a, b F12K066a, b -> F12K071a, b F12K067 -> F12K072 F12K068a, b -> F12K073a, b F12K069a, b -> F12K074a, b F12K070a, b, c, d, e -> F12K075a, b, c, d, e F12K071a, b -> F12K076a, b	3.1
3.1.16	f12kind	In Version 3.0 der Daten waren die neuen Items in Frage 77 der Kindermatrix nicht in ihrer Reihenfolge im Haushaltsfragebogen enthalten. Dies wurde nun korrigiert, so dass sich neue Zuordnungen der Items zu Variablen ergeben. Außerdem wurden hier die Labels der Items so geändert, dass sie die vier Gruppen (bis 6, Kita ja/nein und ab 6, Schule ja/nein) widerspiegeln.	3.1

Nummer	Datensatz	Beschreibung	in Version
3.1.17	f12lela	Die Geburtsländer der Eltern (F12L078A2, F12L078B2) waren in der Version 3.0 nicht mit den jeweiligen Codes besetzt. Dieser Fehler wurde behoben, so dass an den Labels das Geburtsland der Eltern abzulesen ist.	3.1
3.1.18	f12lela	Folgende Umbenennungen wurden vorgenommen, um Konsistenz mit den vorherigen Datensätzen (*f10lela*, *f11lela*) herzustellen: F12L001A -> F12L001C (Geschlecht) F12L001B -> F12L001A (Geburtsjahr) F12L001C -> F12L001B (Geburtsmonat)	3.1
3.1.19	$p, $pgen	Durch Veränderungen in den Berufsverkodungen durch TNS Infratest verändern sich die entsprechenden Variablen (ISCO, KLAS, NACE, etc.) in $p und $pgen. Dies betrifft insgesamt nur wenige Fälle.	3.1
3.1.20	f12p	Bereits in der gleichen Welle im Lebenslauf erhobene Variablen werden im Personenfragebogen überfiltert. Bisher wurde bei betroffenen Personen der Missing-Code "-6" vergeben und die entsprechende Information lag nur in *f12lela* vor. Seit Version 3.1 werden diese Informationen zusätzlich direkt in den entsprechenden Variablen des *f12p*-Files abgelegt. Änderungen treten dabei in 297 Fällen auf.	3.1
3.1.21	f12p	Die Berufsverkodungen (F12P023_IS88, F12P023_KLAS und F12P028_NACE), die vorher nur im *f12pgen* verfügbar waren, wurden in der Version 3.1 hinzugefügt.	3.1
3.1.22	ppfad	Bei 168 Fällen wurde in der Weitergabe 3.0 bei der Variable GEBMONAT nicht auf die neueste Information (aus 2012) zurückgegriffen. Dies wurde in dieser Weitergabe korrigiert und betrifft auch 147 Fälle in der Variable GEBMOVAL.	3.1

Nummer	Datensatz	Beschreibung	in Version
3.1.23	ppfad	Die Migrationsvariablen MIGBACK und GERMBORN wurden überarbeitet, was zu Änderungen in auch in MIGINFO, IMMIYEAR und CORIGIN in insgesamt 237 Fällen führt. Der Großteil der Änderungen betrifft dabei MIGBACK und MIGINFO. 1. Bei GERMBORN und MIGBACK werden für Kinder, die bei Ihren Großeltern leben, nicht mehr die Informationen der Großeltern als Proxy herangezogen. Die Kinder erhalten nun eine "-1". 2. Bei MIGBACK wird nun die Information des einen (deutschen) Elternteils als Proxy für die Kinder verwendet, wenn der andere (deutsche) Elternteil nicht an der Befragung teilnimmt. Dies reduziert den Anteil der Fälle mit "-1". 3. Bei MIGBACK wurde außerdem ein Fehler korrigiert: im Panel geborene Kinder mit nur einem leiblichen Elternteil wurden bisher bei der Generierung nicht erfasst und standen auf "-1".	3.1
3.1.24	$pgen	Aufgrund von Änderungen in der Berechnung von COUPST und MARRST11 kommt es in allen Jahren zu leichten Veränderungen. (Die Variable COUPID wird neu vergeben und ist entsprechend nicht notwendigerweise identisch über die Weitergaben.)	3.1
3.1.25	$pgen	Im SOEP werden seit 2011 neue Bildungsvariablen weitergegeben, hierbei handelt es sich um: FIELD$$ (Fachrichtung) DEGREE$$ (Hochschulabschluss) TRAINA$$ (Ausbildungsberuf, Lehre) TRAINB$$ (Ausbildungsberuf, Berufsfachschule) TRAINC$$ (Ausbildungsberuf, Fachschule) TRAIND$$ (Ausbildungsberuf, Beamtenausbildung) FDT_F$$ (Quelle FIELD, DEGREE, TRAIN) Mit der Weitergabe 3.1 von FiD sind diese Variablen (auch rückwirkend) in FiD enthalten. (Die Variablenlabels sind dabei in deutscher Sprache.)	3.1

Nummer	Datensatz	Beschreibung	in Version
3.1.26	$paradata	Die Datensätze $paradata werden ab dieser Weitergabe nur noch im long-Format als ***paradatal*** weitergegeben. Dadurch fallen die Wellenkürzel (f10, f11, f12) bei den Variablen weg und die Variable SYEAR (Befragungsjahr) wird aufgenommen. Außerdem wird der Datensatz um zwei Variablen erweitert: REQD, die Anzahl der zu beantwortenden Fragen und MISS, die Anzahl der verweigerten oder nicht-validen Antworten, jeweils auf Interviewebene.	3.1

Nummer	Datensatz	Beschreibung	in Version
4.0.1	f13eltern1-3	Die Frage, ob das Kind regelmäßig oder zeitweise durch eine Tagesmutter oder KiTa betreut wird, wurde um die Anzahl der Stunden pro Tag erweitert (F13E1012, F13E218, F13E318). Dadurch wurde in der späteren Abfrage nach der Betreuungssituation auf Tagesmutter und KiTa verzichtet (F13E129, F13E225, F13E325). Siehe auch Punkt 4.0.6.	**4.0**
4.0.2	f13eltern1-4	Es wurden hier neue Fragen eingeführt, in denen die Eltern die Eingewöhnungsphase ihrer Kinder in die KiTa bewerten sollen. Dies führt auch zu neuen Variablen in ***bioage01***, ***bioage02***, ***bioage03***, und ***bioage06*** (siehe Punkt 4.0.7).	4.0
4.0.3	f13eltern2-3	Die Fragen zur Geburt des Kindes, die in den vorherigen Elternfragebögen 2 und 3 noch gestellt wurden, werden für 2013 nicht mehr erhoben - für fast alle Kinder liegt diese Information aus den vorherigen Jahren vor. Siehe auch Punkt 4.0.8.	4.0
4.0.4	f13eltern2-4	Die Aktivitäten, die mit dem Kind gemacht werden (F13E227, F13E327, F13E412), wurden um den Skalenpunkt "4 seltener" erweitert, um die bisherige Lücke zwischen "einmal die Woche" und "gar nicht" zu überbrücken. Dadurch verschiebt sich in dieser Frage abweichend zu vorher "gar nicht" auf den Wert "5". Siehe auch Punkt 4.0.9.	4.0
4.0.5	f13eltern5-6	Es wurden hier neue Fragen eingeführt, in denen die Eltern die Eingewöhnungsphase ihrer Kinder in die Schule bewerten sollen. Dies führt auch zu neuen Variablen in ***bioage08*** und ***bioage10*** (siehe Punkt 4.0.10).	4.0
4.0.6	bioage01, bioage02, bioage03	Durch die Umstellung der Frage nach der Betreuung durch KiTa und Tagesmutter in 2013 (siehe Punkt 4.0.1) ändert sich auch die Kodierung für CARE6, CARE6H, CARE8, CARE8H. Hierbei werden die Angaben zur Betreuung pro Tag auf die Woche umgerechnet. Näheres dazu in der Dokumentation zu den ***bioage*** files (6_bioage01-10.pdf).	4.0
4.0.7	bioage01, bioage02, bioage03, bioage06	Erstmals wurde über die Probleme bei der Eingewöhnung in die KiTa gefragt, diese Variablen heißen ADPCCAR1-ADPCCAR4. Sie sind ausschließlich im Jahr 2013 besetzt und haben sonst den Wert "-5".	4.0

Nummer	Datensatz	Beschreibung	in Version
4.0.8	bioage02, bioage03	Da die Information zur Geburt des Kindes ab 2013 nicht mehr in den jeweiligen Elternfragebögen erhoben wird, werden die entsprechenden Variablen für dieses Jahr auf "-5 Question not included" gesetzt. (Die Information liegt in den meisten Fällen aus den vorherigen Jahren vor.)	4.0
4.0.9	bioage02, bioage03, bioage04	In den Variablen ACTIV1-ACTIV14 wurden durch die Änderungen in f13eltern2-4 (siehe Punkt 4.0.4) ebenfalls die Skalen geändert. Der Wert 4 (bis Version 3.1 "Never") wurde nun zu 5 umkodiert. Neu eingeführt wurde der Wert "4 Less often", der nur in 2013 besetzt ist.	4.0
4.0.10	bioage08, bioage10	Erstmals wurde über die Probleme bei der Eingewöhnung in die Schule gefragt, diese Variablen heißen ADPSCL1-ADPSCL5. Sie sind ausschließlich im Jahr 2013 besetzt und haben sonst den Wert "-5".	4.0
4.0.11	$jugend	Die in den Jugendfragebogen abgefragten Klartexte zum eigenen Berufswunsch sowie zu den letzten Berufen der Eltern werden nun als ISCO-Codes und Klassifizierung des Statistischen Bundesamts bereitgestellt.	4.0
4.0.12	f13jugend	In Frage 15 wurde ein neues Item eingeführt (Nutzen sozialer Online-Netzwerke). Um die Längsschnittkonsistenz zu wahren, wird dieses Item entgegen den Regeln für die Variablenbenennung als "F13J015Q" eingefügt.	4.0
4.0.13	$lela	Die im 2. Teil des Biographiefragebogens als Klartexte erfassten Klassifikationen des ersten Jobs bzw. der letzten Branche werden nun kodiert weitergegeben.	4.0
4.0.15	$mipinc / $mihinc	Neu berechnete Imputationen, die nun auch die in 2013 erhobenen Daten berücksichtigen. Mit den neu berechneten Imputationen ergeben sich auch neue Werte für "LABGRO$" und "LABNET$" in ***$pgen***, sowie neue Werte für imputierte Variablen in ***$hgen***, so dass ausschließlich die neue Distribution genutzt werden sollte. Es gibt außerdem leichte Veränderungen in der Art der Imputation; genaueres findet sich in der entsprechenden Dokumentation.	4.0
4.0.16	f11hgen, f12hgen	Die Variable MOVEYR wurde neu kodiert, nachdem vorher Umzüge von einigen Personen nicht korrekt berücksichtigt worden waren.	4.0

Nummer	Datensatz	Beschreibung	in Version
4.0.17	ppfad	Die Nettocodes in $NETTO wurden überarbeitet und wie im SOEP ergänzt: 180 (Person ohne aktuelle Angabe ohne Austritt), 181 (Vormals Befragte ohne aktuelle Angabe), 188 (Rückkehrer, zuvor Ausland), 189 (Rückkehrer, zuvor Ausfall) sind nun besetzt.	4.0
4.0.18	ppfad	Die Variable LOC1989 (Wohnort Ost oder West 1989) ist nun auf "-1 Keine Angabe" gesetzt, wenn der entsprechende Teil der Biographie nie erfragt wurde und das Geburtsjahr der Person vor 1989 liegt. Vorher war hier eine "-2 Trifft nicht zu" kodiert worden.	4.0