FDZ

Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Datenreport

07/2013
EN

Documentation of labour market data

User Guide „Panel Study Labour Market and Social Security“ (PASS) Wave 6

Arne Bethmann,
Benjamin Fuchs,
Anja Wurdack

Bundesagentur für Arbeit

User Guide "Panel Study Labour Market and Social Security" (PASS)

Wave 6

Arne Bethmann
Benjamin Fuchs
Anja Wurdack (Eds.)

Die FDZ-Datenreporte beschreiben die Daten des FDZ im Detail. Diese Reihe hat somit eine doppelte Funktion: zum einen stellen Nutzerinnen und Nutzer fest, ob die angebotenen Daten für das Forschungsvorhaben geeignet sind, zum anderen dienen sie zur Vorbereitung der Auswertungen.

FDZ-Datenreporte (FDZ data reports) describe FDZ data in detail. As a result, this series has a dual function: on the one hand, users can ascertain whether the data are suitable for their research task; on the other, the reports can be used to prepare the analyses.

Contents

List of Tables

List of Figures

Part I

Introduction to the PASS data

1 Getting started with PASS

Arne Bethmann, Benjamin Fuchs and Anja Wurdack

This User Guide is meant to give information on general issues of the panel study "Labour Market and Social Security" (PASS) and to offer assistance for the work with the datasets of the scientific use file (*SUF*). While the data reports, which are released for every wave, inform in detail about key statistics, data editing, generated variables and the weighting of a certain wave, the User Guide offers comprehensive information that is not specific for a single wave.

Part I gives an introduction to the PASS data, starting in chapter 1 with a first overview of the topics covered by the User Guide and the other working tools that will help users to work with PASS. Information on how to obtain the data can be found there as well. Subsequently, the main research questions which influenced the development of the study will be presented in chapter 2, and it will be pointed out which addition to existing data is made by PASS. In chapter 3, the design, sampling procedure and several special characteristics of the survey design will be described. Chapter 4 deals with the topics of the survey and gives an overview of the subjects of the household and personal interview since wave 1. Thereafter, the structure of the *SUF* and the datasets included will be presented in chapter 5. Not only does this chapter give essential information on the levels, types and formats of the datasets in the *SUF* but also on their topics, key variables and special characteristics. After this overview of the *SUF* and its datasets, chapter 6 focuses on the types of variables that can be found in these datasets and their naming conventions. Subsequently, the general logic of data editing and its most important steps will be discussed. Herein, the standardised missing value codes and special codes that are used in all datasets of PASS are described as part of the section on filter checks (chapter 7). Chapter 8 provides information on the weighting concept, e. g. on the creation of the design weights, the weighting datasets and the variables included.

The second part of the User Guide provides advice and hands-on examples for working with the data. Chapter 9 discusses the use of key variables when merging information from different datasets. The specific handling of the register and spell datasets is introduced in chapters 10 and 11 respectively. An in-depth treatment of the use of weights in cross-sectional and longitudinal analyses is given in chapter 12. The last two chapters discuss specific issues when working with generated variables (chapter 13) and constant characteristics (chapter 14). This part is particularly helpful for new users. It demonstrates certain standard procedures of the work with the PASS datasets.

The User Guide will evolve over time, as it is planed that new topics will be included and already included chapters will be updated in future waves. For this process, feedback from the users of PASS is essential as it can give evidence where the User Guide should go into more detail, which new topics should be considered and where a chapter should be revised or updated. Therefore, we appreciate any feedback, be it positive or negative.[1]

[1] Feedback can be addressed directly via E-Mail to: iab.fdz@iab.de

1.1 The user guides and other working tools

In addition to this User Guide, several other working tools provide information about PASS and its *SUF*. Table 1 gives an overview of the working tools that are currently available via download from the Homepage of the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)[2] and its contents.

1.2 Data access

Currently, the six waves of PASS are available as weakly anonymised *SUF*. The last version of the *SUF* includes information on all waves that have been released before, e. g. the *SUF* of wave 6 includes all information from wave 1 to 6 as well. Wave 7 is expected to be available in autumn 2014.

The *SUF* can be used by researchers at scientific institutions for non-commercial research. Data access is provided by the FDZ of the BA at the IAB. The homepage of the FDZ offers further information on requirements and how to apply for the data.[3]

[2] http://fdz.iab.de/en/FDZ_Individual_Data/PASS/Working_Tools.aspx

[3] http://fdz.iab.de/en/FDZ_Data_Access/FDZ_Scientific_Use_Files.aspx

Table 1: Overview of the working tools available in wave 3

Name	Content	Language	Waves covered
User Guide	The User Guide offers general information on PASS that is not specific to certain waves. The following topics are covered: Objectives and research questions of PASS; Additions to existing data; Survey and Sampling Design; Instruments and interview programme; Structure of the scientific use file and its datasets; General logic of data editing; Weighting concept; Examples on how to use the datasets	English	1–6 (integrated)
Data Reports	For each wave, the respective data report provides wave-specific information on the data editing and tabulations of the surveyed variables in the different datasets of the scientific use file. (Because the user guide was first introduced in wave 3, the data reports of wave 1 and 2 include some of the user guides general information as well.) The following wave-specific topics are covered: Key statistics; Generated variables; Data editing; Weighting; Tabulations of the surveyed variables	English (excluding the tabulations of the surveyed variables)/ German	1–6 (wave-specific)
Methods and Field Reports	For each wave, the methods and field report describes the work of the field institute for the respective wave. The following wave-specific topics are covered: Objectives and design of PASS; Pretest; Detailed information on the steps of the field work; Data editing by the field institute; Weighting (modeling of non-response)	German	1–6 (wave-specific)
Questionnaires	For each wave, the different questionnaires document which items have been surveyed in the respective wave. Furthermore, they make transparent in which variables the collected answers to the items can be found in the scientific use file. Therefore, they establish the correspondence between question numbers and constant variables names. The following questionnaires are available: Household questionnaire for new and split households; household questionnaire for panel households (first introduced in wave 2); person's questionnaire; senior citizens' questionnaire	English/ German	1–6 (wave-specific)

2 PASS background

Mark Trappmann

2.1 Objectives and research questions of the panel study "Labour Market and Social Security"

The panel study "Labour Market and Social Security" (PASS), established by the Institute for Employment Research (IAB), is a dataset for labour market, welfare state and poverty research in Germany, creating an empirical basis for the scientific community and for policy advice.

The study is carried out as part of the IAB's research into the German Social Code Book II[4] (SGB II). The IAB has the statutory mandate to study the effects of benefits and services under SGB II aimed at integration into the labour market and subsistence benefits. However, due to its complex sample design, the study also enables researchers to answer questions far beyond this scope. Five core questions influenced the development of the new study, which are detailed in Achatz/Hirseland/Promberger (2007):

1. Which pathways lead out of receipt of Unemployment Benefit II (UB II)? Which factors facilitate or impede those exits and how do former recipients gain subsistence after having overcome UB II receipt?
2. How does the social situation of a household change when it receives benefits? Apart from the financial situation and the standard of living, the impact on health or social exclusion is of interest here.
3. How do the individuals concerned cope with their situation? Is there a change of attitudes or behaviour over time?
4. In what form does contact between benefit recipients and institutions providing basic social security take place? What are the actual institutional procedures applied in practice?
5. What employment history patterns or household dynamics lead to receipt of UB II?

2.2 How does PASS fit in the German microdata landscape?

When PASS was designed, German labour market, poverty and welfare state research already had access to various excellent micro-data. In particular, there were a number of longitudinal datasets available which already covered relatively long survey periods. A particularly important survey data source is the German Socio-Economic Panel Study (SOEP) (Wagner/Frick/Schupp, 2007), which provides annual data at the individual and household level dating back to 1984. In addition, administrative data from the Federal

[4] Social Code Book II – Basic Social Security for Jobseekers (Sozialgesetzbuch (SGB) Zweites Buch (II) – Grundsicherung für Arbeitsuchende).

Employment Agency (BA) is processed at the IAB and provided for research use by the Research Data Centre (FDZ) of the BA at the IAB, for example in the form of the *Integrated Employment Biographies (IEBS)*, the *IAB Employment Samples (IABS)* or the *Linked Employer-Employee Dataset (LIAB)*.

The spectrum of questions and the design of PASS are intended to close gaps in the existing stock of data. PASS has three main characteristics that extend analysis potential beyond that of the Federal Employment Agency's administrative data:

1. The panel takes the household context into account – including the situation before and after receipt of UB II.
2. The panel is complete in that it covers all groups of persons and all employment biographies, not only people in dependent employment, unemployed people and those in need of assistance. The dataset also provides information on the status during phases of economic inactivity, self-employment or employment as civil servants.
3. The panel collects additional or significantly more detailed data on relevant characteristics such as attitudes, employment potential or job-search behaviour.

Compared to the existing surveys of individuals or households, PASS aims to improve the data situation in particular with regard to the following points:

1. The high case numbers of UB II recipients (cf. section 3) make it possible to conduct more detailed analyses – for example on the impact of SGB II on certain target groups like young adults, migrants, single parents, supplemental benefit recipients ("Aufstocker") – and to obtain more precise estimates of statistics and model coefficients than from datasets in which benefit recipients are only included in proportion to their share of the population.
2. Collecting additional characteristics such as the intensity and type of contact to institutions providing basic social security or participation in employment and training measures makes it possible to analyse the significance of institutional assistance for the population below the poverty line.

Linking the survey data with the administrative data of the BA enables validating the characteristics surveyed and also conducting analyses in which the higher measurement precision of the process-generated data can be combined with further variables and the household context from the survey.

3 Design of the study

Mark Trappmann, Gerrit Müller and Arne Bethmann

3.1 Introduction

By establishing the panel study “Labour Market and Social Security” (PASS), the Institute for Employment Research (IAB) is setting up a database that creates a new empirical basis for research into the labour market, the welfare state and poverty in Germany. The survey pays particular attention to the dynamics of households in receipt of benefits in accordance with the Social Code Book II (SGB II) (see chapter 2.1 on the objectives and questions of PASS, and in more detail Achatz/Hirseland/Promberger 2007: 17 pp.).

An adequate survey design has to be tailored to the research demands and the population to be surveyed. The strategies employed in PASS are described in subsection 3.3. They are further detailed in Schnell (2007) and Rudolph/Trappmann (2007).

The most important decisions that were taken in PASS are those for a prospective longitudinal design and for conducting it as a household survey. The main research questions require longitudinal data: They ask for determinants of inflows into and outflows from benefit receipt or for changes in attitudes, action taken or the material situation before and after the beginning of benefit receipt. The only adequate design to answer such questions is the panel design where the same units of observation are asked to answer the same questions in repeated waves. In PASS the period of time between two consecutive waves was – based on expectations on how quickly important target variables change – devised to be one year.

When examining research questions in the context of the SGB II the respondents’ action context and in particular here their household context is of importance for two different reasons: First, because individuals make decisions against the background of their household-specific circumstances. Second, because the SGB II also examines the household context when activating benefit recipients, in the context of “support and demand” (see Achatz/Hirseland/Promberger 2007). Therefore, PASS is designed as a household survey: within a household all members aged 15 or above are to be interviewed with a person-level questionnaire. The personal interviews are always preceded by a household interview in which general, household-related information is gathered.

In section 3.3 the reader will find information on the sampling design, while section 3.3 contains other design aspects like mode, interview languages, interviewer trainings etc.

3.2 Sampling procedure

The two main features of the sampling design are the dual frame (Unemployment Benefit II recipients (UB II recipients) and general population) and the yearly refreshment of the UB II sample by new entries to the population:

- Analyses of inflows into receipt of UB II, comparisons of households in receipt of benefits with households not receiving benefits, the investigation into hidden poverty and the formation of control groups require a comparison of benefit recipients with the rest of the population. For this reason PASS combines a sample of benefit recipient households with a sample of the general population disproportionately stratified according to status.

- In order to be able to analyse inflows into receipt of UB II already after a short time and to guarantee the representativeness of the sample of benefit recipient households in the cross-section, a refreshment sample for this group is drawn in every wave (on the concept of the refreshment sample see Trappmann et al. 2009: 11 pp.).

Therefore, the sample in the 1st wave of PASS consisted of two subsamples. These two otherwise independent samples are connected on the first sampling stage via the selection of identical primary sampling units (for detailed information about the sampling design of the 1st wave see Rudolph/Trappmann 2007: 65 pp.). The first subsample (BA sample) is a random sample of benefit units ("Bedarfsgemeinschaften"), in which at least one person was receiving UB II in July 2006. This sample was drawn from administrative data of the federal employment agency (BA). As PASS is a household survey, the entire household in which a benefit unit was living was targeted for the survey. The second subsample is a sample of private households in Germany (general population sample). In wave 1 a random sample of addresses was drawn from the *MOSAIC database* of addresses held by the commercial provider Microm. The sample was stratified disproportionately by status in such a way that households with a low social status and thus a greater risk of entry into benefit receipt had a higher probability of inclusion (on the results of the stratification see Trappmann et al. 2007).

On the first sampling stage 300 postcodes were drawn from the postcode register. These postcodes serve as primary sampling units in PASS (on the selection of the primary sampling units see Rudolph/Trappmann 2007: 77 pp.). The selection probability of a postcode sector was dependent on the number of households in the particular sector according to the *MOSAIC database* (probability proportional to size). Within each sampling point, benefit units (BA sample) or addresses (general population sample) were drawn. The number of benefit units to be drawn for the BA sample depended on the rate of benefit recipients (number of benefit units in the sampling point according to BA process data divided by the number of households in the sampling point according to the *MOSAIC database*). On average 100 benefit units were selected into the gross sample per sampling point. As the number of selected benefit units is proportional to the benefit recipient rate in the sampling points, a uniform selection probability is also guaranteed in the BA sample (Rudolph/Trappmann, 2007: 78 pp.). All members of each household in which a benefit unit was living were surveyed. In the initial wave 6,804 households from this sample were interviewed.

For the general population sample 100 addresses were drawn within each sampling point. In order to obtain an overrepresentation of the lower status classes, addresses of lower status classes had a higher inclusion probability. The addresses drawn in this way were visited by employees of the field institute conducting the survey, who copied all names on the doorbell panels. At the field institute a random selection of these doorbells was made. If a doorbell panel had more than one name on it, one of these names was selected. Each selected person's entire household was surveyed. In the initial wave 5,990 households from this sample were interviewed.

All households in the two samples of the 1st wave were to be re-interviewed in the 2nd and all consecutive waves (see the corresponding data report for response rates, e. g. Berg et al. 2013c for the 6th wave). In addition to this, households that had split off from the households interviewed in one of the preceding waves were also surveyed. They were each assigned to the subsample from which their original household had been drawn (either of the two subsamples in the 1st wave or a refreshment sample in one of the later waves).

In addition, starting with the 2nd wave, for each wave a refreshment sample was drawn from benefit units that had begun receipt of UB II. These are benefit units which were receiving UB II at the refrence date for sampling of a specific wave (i. e. July of the year prior to the respective survey wave, e. g. 2007 for the 2nd wave up to July 2011 for the 6th) but not at the sampling date of all preceding waves. The sample was drawn within the primary sampling units of the 1st wave following the procedure used in the first wave. The size of these yearly refreshments is about 1,000 households. The aggregate of all households from all BA subsamples who still receive benefits at the most recent reference date can be projected to all households with at least one recipient of Unemployment Benefit II in Germany at that time.

In the course of a panel former respondents attrite, reducing the number of units (households, persons) left for analysis. To counteract this process PASS replenished both samples before wave 5. To distinguish these additional samples from the refreshment samples above which focus on new entries to the population, we will refer to them as replenishment samples. 100 new primary sampling units were selected for this replenishment (again with probability proportional to size).

This time the general population sample was however drawn from municipal registers without stratification. Details of the procedure can be found in chapter 6 of Berg et al. (2013b). The Replesnishment of the BA sample was drawn according to the same procedure as all previous BA samples. The population of the refreshment was designed to be all households with at least one benfit unit in receipt at the wave 5 reference date for sampling (July 2010).

The Methodology used to estimate weights that allow for an integrated analysis of the ongoing panel samples with the replenishments was described in Spiess/Rendtel (2000). For details on weighting the reader is referred to chapter 8 of this User Guide. In the PASS dataset all subsamples can be identified by the variable `sample` in the household dataset (*HHENDDAT*). It uses the following values:

1: wave 1 BA sample

2: wave 1 general population sample (from Microm addresses)

3: wave 2 refreshment of BA sample (new entries)

4: wave 3 refreshment of BA sample (new entries)

5: wave 4 refreshment of BA sample (new entries)

6: wave 5 general population replenishment (from municipal registers)

7: wave 5 replenishment of BA sample (from municipal registers)

8: wave 5 refreshment of BA sample (new entries)

9: wave 6 refreshment of BA sample (new entries)

3.3 Other survey design features

PASS is administered to a particularly difficult survey population that is usually underrepresented in surveys. A substantial part of the sample consists of benefit recipients and low-income households with, on average, a rather poor level of formal education and low social status. A number of survey design characteristics and fieldwork procedures have been adopted to reduce initial nonresponse and panel attrition as well as selectivity of nonresponse with respect to important target variables.

3.3.1 (Sequential) Mixed-mode design

PASS uses a mix of computer-assisted telephone interviews (CATI) and computer-assisted personal interviews (CAPI), with CATI as the default mode in waves 1–3. The mixed-mode design was chosen as a cost effective way of addressing various issues related to low income and welfare populations (Rudolph/Trappmann, 2007: 91–92). Particular problems faced when trying to interview these groups are, for example, their tendency to relocate more frequently than the general population, difficulties in contacting them by phone due to low landline coverage, or changes in mobile phone numbers. The sequential mixed-mode design ensures that target persons who cannot be contacted and interviewed by phone, are visited by an interviewer at their home to conduct the interview in CAPI mode. Initiated by a call for tenders, PASS changed the fieldwork agency responsible for data collection and preparation after wave 3 (see Müller, 2011). In the course of that change there has also been a shift towards CAPI as the default mode for cases of all refreshment samples from wave 4 onwards. For details on contact routines and mode switches please see subsection 3.3.3.

3.3.2 Foreign-language interviews

The design anticipates that a considerable proportion of the target population has a migrant background and may not have sufficient knowledge of German to participate. Therefore, the survey instrument was translated into Turkish and Russian, the most frequent first languages of immigrants to Germany. In wave 1, there was an additional English-language version as a fall-back for all other nonnative speakers. Since only a small number of cases was realised using the English version of the instrument (9 household-level interviews), it was dropped after wave 1. In the CATI telephone survey, the foreign-language instruments were administered by interviewers who were native speakers of the respective language. As a cost saving measure, the strategy employed in CAPI mode was to transfer respondents back to the telephone field whenever possible. Where this could not be done, the CAPI interviewers used a written foreign language version of the respective questionnaire as translation aid. For wave-specific information, see Hartmann et al. (2008: 19–20) (wave 1), Büngeler et al. (2009: 12–14) (wave 2), Büngeler et al. (2010: 17) (wave 3). These procedures remained largely unchanged in waves 4–6, except that no written translation aides were used anymore in combination with a German-language version in face-to-face interviewing. All foreign-language interviews were done by native speakers of the respective language, with a fully translated survey instrument. In practice, as before, most foreign-language interviews were conducted in CATI. For details, see Jesske/Quandt (2011: 19, 25, 44–46) (wave 4), Jesske/Schulz (2012: 29, 37, 57–61) (wave 5), Jesske/Schulz (2013: 25, 34, 58–60) (wave 6).

3.3.3 Fieldwork procedures: contact routines, mode switches and refusal conversion

By default, in waves 1–3 contact was first attempted by telephone whenever a number was known to exist for a particular address, either because it was part of the information on the sample frame, or because it could be traced by phone number search prior to the beginning of fieldwork. Cases for which no valid telephone number was available, started off in CAPI mode. As noted above, from wave 4 onwards there has been a shift towards CAPI as the default mode for cases of all refreshment samples, with panel cases being first contacted in the mode of the previous interview. Across all waves, a mode switch from CATI to CAPI took place if at least twelve consecutive contact attempts by telephone were unsuccessful, or if the household explicitly asked for being interviewed face-to-face. Similarly, cases could also be switched from CAPI to CATI mode. This happened automatically if six consecutive contact attempts were unsuccessful or if a household requested to be interviewed by phone. Contact attempts in both survey modes were varied across weekdays and daytimes in order to minimise household nonresponse due to noncontact. For further details on the organisation of fieldwork in each wave, please see the survey agencies' field reports; Hartmann et al. (2008: 20–44), Büngeler et al. (2009: 14–29), Büngeler et al. (2010: 22–40) (waves 1–3), and Jesske/Quandt (2011: 23–28), Jesske/Schulz (2012: 36–41), Jesske/Schulz (2013: 33–38) (waves 4–6). Note that in waves 1–3 the interview mode was determined at the household level, that is all respondents within a given household were interviewed in the same mode. As of wave 4, this could be handled more flexibly with interview mode being assigned individually, that is at the person level. In each wave, there were refusal conversion attempts by telephone towards the end of the fieldwork period for households or persons who initially refused to participate for the following reasons: lack of interest in the topic, length of the interview, lack of time, when someone immediately hung up the phone, or when someone that was not the target respondent refused on behalf. This follow-up on reluctant/hard-to-interview sample cases was conducted by selected CATI interviewers with above average performance during the regular fieldwork and special training in refusal conversion (see e. g. Hartmann et al. 2008: 54–56 (for waves 1–3), and e. g. Jesske/Quandt (2011: 26) (for waves 4–6)). Indicator variables for interview mode are contained in the PASS scientific use file in order to control for potential mode effects in empirical analyses.

3.3.4 Advance letter and other survey notification material

In wave 1, each household in the gross sample was notified with an advance letter about upcoming calls or personal visits by interviewers approximately one week prior to the first scheduled contact attempt. The letter introduced the name and purpose of the survey, the involved research institutes (IAB, fieldwork agencies), and the sponsor (Federal Ministry of Labour and Social Affairs). It explained how the respective household was selected into the sample and that all data protection laws would be strictly adhered to. Respondents were given a promise of confidentiality which guaranteed that their names and addresses would be kept separately from any of the information they provided in the survey and would not be passed on to third parties. The letters were tailored to the two subsamples (register vs. population sample), stressing the importance of response to the survey request, yet emphasizing that participation was voluntary (Hartmann et al., 2008: 43, 78–83). From wave 2 onwards, additional versions of the advance letter were developed, tailored to panel households that had already participated in the previous wave(s); see e. g. (Büngeler et al., 2009: 29, 62–65) (for waves 2–3), and e. g. Jesske/Quandt (2011: 90–134) (for waves 4–6). New entrants to the study, such as cases of the yearly refreshment samples, received a revised version of primary notification letter. In all waves, a thank-you letter was mailed out to each respondent shortly after the interview in order to increase the propensity to participate in future waves. In addition, a newsletter was mailed out to respondents between waves providing them with some results from prior waves with the main objective to build rapport with respondents through means other than the annual interview itself.

3.3.5 Tracking

One of the top priorities in an ongoing panel survey is to maintain up-to-date and accurate records of the whereabouts of each sample member. In PASS, both prospective/proactive and retrospective tracking procedures (Couper/Ofstedal, 2009; Laurie/Smith/Scott, 1999) are being used in conjunction. Prior to the beginning of each wave's fieldwork, attempts are made to update address and contact information. In PASS, this happens primarily on the basis of the thank-you letter mailing to previous wave's respondents and the mail-out of advance letters to all sample members of the current wave (i. e. including temporary dropouts and newly issued cases from refreshment samples). In both instances, the returned mail identifies addresses with need for tracking prior to the beginning of the actual fieldwork. In wave 1 and 2, movers were attempted to be traced through address information provided by the *Deutsche Post* on the return mail or by a request to the residents' registration office (Einwohnermeldeämter) at a household's last known address. As of wave 3, additional resources have been committed to tracking. First, a specialised tracking service of *Deutsche Post* called "*Addressfactory*" was used as a supplementary source to update and search for addresses. Second, an additional update of address information and telephone numbers was conducted on the basis of administrative records available at the BA before the respective wave's fieldwork period. The proactive tracking procedures just described remained largely the same in waves 4–6, i. e. after the change in fieldwork agency. Retrospective tracking sets in during fieldwork when interviewers discover that a sample member does not live at the designated address (anymore) or the telephone number is not/no longer valid. In CAPI mode, interviewers would try to obtain address and phone information from neighbours or present occupants at the respondent's former address. If unsuccessful, these cases are forwarded to the centralised tracking system and searched for in the various databases and registers described above. The same procedure is applied to CATI cases with invalid telephone numbers. In waves 1–3, centralised tracking was not (yet) performed on a continuous basis for each individual address mover, but only at a few designated points during the fieldwork period in "batches" of addresses. However, tracking efforts during fieldwork were continuously intensified. From wave 1 to 3, the number of time points at which searches via residents' registration offices were initiated was increased from three to five. For details on the tracking procedures in each of the first three waves, please see Hartmann et al. (2008: 22–23, 31–33) (wave 1), Büngeler et al. (2009: 15–16, 20–22) (wave 2), Büngeler et al. (2010: 24–25, 28–33) (wave 3). As of wave 4, tracking happened on an almost continuous basis and using all search channels (e. g. residents' registration offices, "Addressfactory" of Deutsche Post, telephone registers, administrative records, as well as information from tracking efforts by face-to-face interviewers) concurrently. For further details on the tracking procedures as of wave 4, see Jesske/Quandt (2011: 28–41) (wave 4), Jesske/Schulz (2012: 41–54) (wave 5), Jesske/Schulz (2013: 38–54) (wave 6).

3.3.6 Respondent incentives

As in many other household surveys, PASS distributes incentives for respondents in order to increase response rates and potentially bound the scope for nonresponse and attrition bias. In wave 1, all sampled households received a special postage stamp, as a small token of appreciation, together with the advance letter. In the advance letter it was stated that respondents to the survey would receive a ticket for the lottery *"Aktion Mensch"*. The ticket had a value of about 1,50 EUR and was mailed to each individual respondent after the interview together with the thank-you letter. In wave 2, the type of incentive strategy was left unaltered, with the exception that the ticket was now for the lottery *"ARD-Fernsehlotterie"* and had increased in value to about 5,00 EUR. Flanking the other measures to increase survey participation as described (e. g. extended field period, increased tracking efforts), in wave 3 there was also a shift in incentive strategy towards the usage of monetary incentives. A new incentive scheme was introduced that consisted of a 10,00 EUR note distributed at the household-level in advance of the interview, i. e. unconditional on participation. It was sent to each panel household that had participated at least once together with the advance letter. A split-sample experimental design was used in order to be able to evaluate the effects on response rates, sample composition and bias afterwards (see Felderer et al., 2012). Households (new entrants) of the wave 3 refreshment sample were not part of the experiment and kept receiving the unchanged incentive, i. e. the lottery ticket conditional upon participation for each responding household member individually (Büngeler et al., 2010: 18–20). In waves 4–6, monetary incentives were maintained and further extended in two ways: First, for panel households, the unconditional monetary incentive (10,00 EUR) was now distributed at the person-level, i. e. to each individual household member eligible for an interview. Second, also cases of the respective wave's refreshment sample(s) started receiving the same monetary incentive after the interview, i. e. conditional on response (Jesske/Quandt, 2011: 31–32). Across all waves, and in addition to incentives distributed centrally by mail, face-to-face interviewers were equipped with "doorstep-incentives" (e. g. small gifts) which they could deploy at discretion in order to gain cooperation.

3.3.7 Interviewer training

Shortly before the beginning of each wave's fieldwork, a one-day (as of wave 4: two-day) intensive training programme was offered to interviewers in order to familiarise them with the specific survey requirements. Only experienced interviewers who had previously worked on comparable studies or who had passed a mandatory two-day general interviewer training by the fieldwork agency were admitted to PASS. The study-specific training program provided an introduction to the survey topic and target population, followed by an overview of the questionnaire modules and some hands-on exercises with the programmed instrument. In waves 1–3, all CATI interviewers were directly trained by IAB researchers and programme directors at the fieldwork agency. For CAPI interviewers the training was organised as a train-the-trainers program ("Multiplikatorenkonzept"). That is, a small group of experienced interviewers ("Kontaktinterviewer") was trained centrally and went out to instruct other interviewers in the various geographic areas/sampling points (for details see Hartmann et al. 2008: 34–37 (wave 1), Büngeler et al. 2009: 22–24 (wave 2), and Büngeler et al. 2010: 33–35 (wave 3)). As of wave 4, the train-the-trainers program was abandoned, and all CATI and CAPI interviewers were trained centrally by IAB researchers and trainers of the responsible fieldwork agency (see Jesske/Quandt 2011: 66–68; Jesske/Schulz 2012: 79–81; Jesske/Schulz 2013: 83–85). In all waves, interviewers could keep the training materials and additionally received an interviewer project manual as a comprehensive reference for later (for an example, see the training material for wave 5 which has been published as FDZ-Methodenreport: Beste et al. 2011). In order to keep survey non-cooperation low, the IAB requires the survey agencies to employ a special training course for interviewers: the "refusal avoidance training" (RAT) by Schnell (Schnell/Dietz, 2006), which is based on Groves/McGonagle (2001). It instructs interviewers how to deal with typical arguments of designated respondents who are reluctant to participate in the survey. Interviewers, who have already participated in a prior wave, receive a somewhat shorter training focussing only on changes to recruitment protocols and instrument changes. Interviewers new to the survey in each wave, always receive the full initial PASS training program as described above.

3.3.8 Sampling frame and auxiliary data for nonresponse analyses and post-survey adjustments

In addition to the survey design characteristics that were adopted to reduce nonresponse and panel attrition ex-ante, an unusually good database is available for nonresponse analyses and post-survey adjustments. First, the population samples were either drawn directly from the database *MOSAIC* by Microm Consumer Marketing (wave 1) or later linked to it using address information (wave 5). MOSAIC contains a number of auxiliary variables available at the address level of a sampled unit that can potentially be used, for example, to predict survey non-cooperation (e. g. by indicators of social status or of privacy concerns), or whether a sampled unit can be localized/contacted successfully (e. g. by the proportion of households moving away from a designated area in the course of a year) and which have been used in post-survey adjustment for the wave 1 population sample. A detailed description of the MOSAIC database can be found in Kueppers (2005). Second, the administrative record data on benefit recipiency used for drawing the register samples offers an even richer database in that regard. It contains information (e. g. level of schooling, age, current employment status) at the individual level that can be used to analyse and correct for initial nonresponse of UB II recipient sample cases (e. g. Schnell et al. 2010).

3.3.9 Record linkage to administrative data

In order to further enhance PASS survey data, individual survey responses have been linked to administrative data of the Federal Employment Agency with additional information on episodes of UB I and II receipt, employment, job search, and participation in active labour market programmes; for respondents who gave their consent to record linkage. For the wave 6 data release, about 87,2 % of the individuals responding to the person-level questionnaire, could be linked successfully (Trappmann et al., 2013). For technical issues regarding the linkage methodology, please see the report by Bachteler (2008). Empirical analyses of the determinants of consent to record linkage and/or potential selectivity biases it may introduce can be found in Beste (2011). In cooperation with the Research Data Center (RDC) of the Federal Employment Agency at the IAB and PASS, a combined survey and administrative dataset (named PASS-ADIAB) comprising all PASS respondents and variables plus administrative information from the SIAB dataset (Dorner et al., 2010) is currently in preparation. This integrated file will be available to researchers for onsite use at the RDC in Nuremberg or one of the RDC locations at Berlin, Bremen, Dresden, Düsseldorf, Mannheim or Ann Arbor (see `http://fdz.iab.de/en/FDZ_Scope_of_Services.aspx` for details and locations). This combined dataset will allow for both, substantive research that treats the administrative data as a supplement with additional information, and methodological research that uses administrative records as a validation source, e. g. in studies of measurement error in survey responses (e. g. Kreuter/Müller/Trappmann 2010).

3.3.10 Call records, fieldwork monitoring and the development towards responsive design

In recent years, survey agencies have started to collect basic call history data on a routine basis. Call history data contain, for example, information about the day and time of the call to a sample unit and the outcome of that call. In an interviewer-administered survey like PASS, with both face-to-face and telephone components, a call is referred to as either a visit to a household or a telephone call, depending on the assigned mode. Survey researchers hope to employ such call records to inform and improve responsive and adaptive survey designs (Groves/Heeringa, 2006; Laflamme/Maydan/Miller, 2008), that is, to monitor the survey data collection process and to respond to changing survey conditions in real time. For example, the available information may be used to identify more difficult cases with regard to establishing contact or cooperation, and to flag sample units that require further follow-ups or a different calling strategy. Call record data may also help to evaluate interviewer performance and to inform calling strategies and interviewer training longer-term. Another advantage is that such data are typically available for both respondents and nonrespondents. In particular in combination with variables from the sampling frame and other auxiliary information, such survey process data ("paradata", Couper 1998; Kreuter 2013) may be of help in nonresponse modeling and in designing effective interventions to reduce nonresponse and nonresponse. In PASS, detailed interviewer call record data have been made available by the respective fieldwork agency only as of wave 3, but have since been delivered on a regular basis throughout any subsequent field period. The data are at the call level and include for any PASS sample member the time and day of each call, the outcome of the call, the mode, whether a call was a regular or a refusal conversion call, or part of the Turkish or Russian language fieldwork. For wave 3, call records are available at the household level and for the telephone field only. As of wave 4, the call record data also contain any call at the person level (individual respondents within households) and from the face-to-face field. Initially, the call record data have mainly been used for purposes of monitoring key indicators (e. g. contact and cooperation rates, number of fixed and vague appointments, nonworked cases, cases with need for tracking) independently of the respective fieldwork agency and for communication with responsible field managers. Since wave 4, information from the call history file has been linked with variables from the administrative sampling frame (for BA refreshment samples) in order to monitor the development of nonresponse bias along a number of dimensions. Since wave 5 this has been further extended to estimating response propensity models repeatedly during the field period and to monitoring the development of the sample composition by means of R-indicators (Schouten/Cobben/Bethlehem 2009, www.risq-project.eu). In wave 6, towards the end of the field period, this information was used to initiate a case prioritization of likely nonrespondents in combination with increased interviewer incentives, similar to Peytchev et al. (2010), in an effort to reduce the risk of nonresponse bias. Wave 6 also saw the first attempt in PASS to actively use the available call record information to enhance call scheduling algorithms, i. e. to assign call attempts to beneficial call windows (daytimes and weekdays) in order to reduce the number of attempts needed to reach a sample unit and to increase cooperation given contact was established (Kreuter/Müller, forthcoming).

4 Instruments and interview programme

Jonas Beste, Johannes Eggs, Stefanie Gundert and Claudia Wenzig

To address the study's key research questions (cf. section 2) the PASS questionnaire covers a broad range of information on individuals and their households. Therefore information is collected by means of separate questionnaires at the household level and the individual level. First, the head of each household answers a household questionnaire. In this interview information referring to the entire household is gathered. In addition, for each household member aged 15 years or older, there is a personal interview about the personal situation of the particular household member. Household members from the age of 65 are interviewed on the basis of a so-called senior citizens' questionnaire. This is a short version of the individual questionnaire and excludes questions that are less relevant for this age group. Below, referring to the main questions of PASS the issues of the different questionnaires are described (see also Trappmann et al. 2013, 2010).

In order to analyse the duration and dynamics of receiving welfare benefits a core part of the survey is composed of data on Unemployment Benefit II (UB II) receipt of the household. The data are gathered retrospectively in the household questionnaire. For newly interviewed households (e. g. of the refreshment sample) information on benefit receipt is collected for a period of about two years before the respective wave. For example, in wave 6 (2012) all episodes since January 2010 were relevant. For re-interviewed households dependent interviewing is used and benefit receipt is always recorded for the period since the last interview date. For each episode, there is information about its beginning and end (month and year), and if applicable, about reasons for and duration of UB II cuts. Furthermore we ask about reasons which lead into and out of receiving benefits (since wave 4).

The household-level information about UB II receipt is complemented by two questionnaire modules on the personal level, which all persons living in households that receive UB II should answer: The first one includes questions about individuals' contact to the agencies responsible for the provision of UB II (e. g. contact frequency, perceived quality and intensity of support). The second module consists of retrospective questions about individual participation in labour market policy programmes. The concept of surveying participation in active labour market programmes (ALMP) was thoroughly reworked in the first waves (for further information, see Gebhardt et al. 2009). Especially, difficulties in identifying clearly the exact type of employment and training measures necessitated the revision. Since the last revision (wave 4) the module exclusively relates to one-euro-jobs (start and end of the programme, evaluation of institutional support during the programme, perceived success of the programme).

Information on unemployment insurance payments (so called UB I) is also collected on the personal level. Questions about the start and end date of the benefit receipt and reasons for ceasing unemployment are integrated in the biography module where the employment history of all participants is gathered. All episodes of employment, unemployment and non-employment (e. g. education, maternity leave, retirement) are recorded, beginning approximately two years before individuals' first interview and are updated in each panel wave. The start and end date (month/year) of each episode is surveyed to ensure a chronological and complete employment history (furthermore parallel activities should be reported and a control mechanism is developed to show gaps of more than one month between the reported activities). For each job episode detailed information is obtained like occupation, job level, type of contract, working hours, income, industrial sector, firm size. Similar characteristics are also collected for the first job ever. For mini-jobs only cross-sectional data is collected (working hours and income of the current mini-job).

With the combination and linkage between the history of UB II receipts and (un)employment history on individual and household level PASS provides excellent data to answer the pathways which lead into and out of UB II receipt and how sustainable and enduring these exits are (see core questions 1 and 5).

Starting in wave 6, information on the "Educational Package" (Bildungs- und Teilhabepaket), which was introduced in 2011 for needy children from low-income families, is collected. The new benefit provides financial means to enable children to take part in sports, music and cultural activities as well as to go on school trips, have lunch in day care centres/schools or receive learning support if they are at risk of having to repeat a school year. Respondents are asked whether they know about the new benefit and, if so, to name the sources of information. Furthermore data is available about the (non-)take-up of the new benefit (separately for each activity), reasons for (non-)take-up and the evaluation of the new benefits. Information on participation in the respective activities is not exclusively collected for children who are eligible for benefits of the Educational Package, but for all children and youths below the age of 25. Thus participation rates can be compared.

The study's second main research question refers to the living conditions of individuals and households. Therefore PASS includes a broad range of questions concerning the households' financial situation, the standard of living (degree of material deprivation), residential conditions, the domestic situation, individual health, care for elderly persons, social networks and social integration. In the following two dimensions should be described in detail: financial situation and health. There are several income measures to assess the material situation of the households ("resource approach"). At the household level there is information on the total disposable household income, saving, loans and debts as wells as security benefits and transfers between households. Individual income components are gathered mainly within the personal questionnaire. They refer to income from work (including earnings from self-employment or special payments), unemployment insurance payments, as well as statutory or civil servants' pensions.

As income measures are prone to short-term fluctuations (e. g. due to changes in household composition or household members' employment status), standard of living is measured in a more comprehensive way by accounting for economic deprivation. Deprivation can be defined as the non-availability of basic goods considered essential for an appropriate standard of living in a society. PASS includes a weighted deprivation index of 26 items, covering basic goods and technical equipment (like having an apartment with separate bathroom or balcony, television or DVD player), activities satisfying basic needs (such as having a hot meal per day, or buying necessary medication), and leisure activities (like going to the movies, or inviting friends for dinner at home once in a while). The index refers to the number of items respondents miss for financial reasons. Each item is weighted by the proportion of respondents regarding it essential.

Health-related questions are gathered by a basic module (in each wave) and by extended questionnaire modules (in selected waves). The basic module includes self-reported health indicators referring to mental and physical well-being: number of consultations, hospital stays in the last twelve months, having chronic diseases and disabilities, mental health problems, subjective health status and health satisfaction. In addition to the basic set, questions about health-related behavior, such as drinking and smoking, obesity and health-related limitations of employment are integrated in selected waves. Furthermore, in this extended module (wave 3 and 6) the standardized SF-12 item battery is used as a global measure of mental and physical health. Besides, new questions about sports activities (types of sports, intensity of physical excercise, social networks related to sports activities, sports activities in youth) were developed for another focus module referring to health (wave 6 and 7).

Referring to the third core dimension of the study (coping strategies, attitudes and behaviour) a variety of subjective indicators are gathered at individual level, such as global aspects of personality traits and situation-specific behavior or attitudes. The basic questionnaire programme includes items on perceived self-efficacy, general life satisfaction, satisfaction with health and living conditions, religiosity and confession, employment orientations as well as the respondents' subjective perception of social integration. The module on job search behavior is another basic element of the questionnaire. Here, the following information is available: search duration, activities to find a new job, reasons for not looking for a job, preferences regarding job attributes (e. g. working hours), willingness to make concessions (e. g. regarding wages, commuting distance, qualification mismatches) and reservation wages. Some subjective indicators are not included in each wave (see table 1), e. g. gender role attitudes, opinions on handling money and partnership or aspirations for children's education and leisure activities. Furthermore, the questionnaire of wave 5 covers the so-called Big Five personality traits, a standardized 21-item battery covering five broad domains of personality.

Finally, PASS includes a broad range of important socio-demographic background variables, like household size and household composition. In addition, questions about child care and school type of children under 15 years are integrated in the household questionnaire. At the personal level questions on education and vocational qualifications are asked, for panel respondents the information is updated every wave. Information on the highest level of education and vocational qualification is available also for respondents' parents. Furthermore, the module on social background covers parents' occupational status and the jobs they held when the respondent her/himself was 15 years old. Migration background is recorded dating back to the third generation. For migrants, additional information on their residential permit is collected. Furthermore, the language spoken in the household and German language skills are surveyed in selected waves.

Table 2 gives a detailed overview over wave-specific modules of the household and personal questionnaires (waves 1–6).

Table 2: Overview of modules

	Wave					
Modules on household level	1	2	3	4	5	6
Household composition	x	x	x	x	x	x
Standard of living (deprivation)	x	x	x	x	x	x
Language spoken in household	x	x	x			
Housing	x	x	x	x	x	x
Receipt of Unemployment Benefit II (UB II)	x	x	x	x	x	x
Income	x	x	x	x	x	x
Child care	x	x	x	x	x	x
Participation of children and youths						x
Education and inclusion subsidies						x
Modules on individual level						
Date of birth	x	x	x	x	x	x
Religion	x	x		x	x	x
Migration	x	x	x	x	x	x
Language spoken in circle of friends	x	x	x	x		
Knowledge of German language				x	x	x
Social origin	x	x	x	x	x	x
Satisfaction with life in general, health and living circumstances	x	x	x	x	x	x
Perceived integration in the society	x	x	x	x	x	x
Self-efficacy beliefs	x	x	x	x		x
Personality: Big Five					x	
Work orientations	x		x	x	x	x
Gender role attitude	x	x			x	
Attitudes towards handling money and partnership		x			x	
Aspiration for children's education	x	x	x			
Education/training	x	x	x	x	x	x
Employment (employment history since January 2005; first/ last job; pooled measures on the entire employment biography) and currently earned income	x	x	x	x	x	x
Unemployment and receipt of Unemployment Benefit I (history since January 2005; pooled measures on the entire unemployment history)	x	x	x	x	x	x
Mini-jobs	x	x	x	x	x	x
Job-search	x	x	x	x	x	x
Participation in employment and training measures	x	x	x			
Participation in 1-euro-jobs				x	x	x
Contact to social security institutions	x	x	x	x	x	x
Perceptions about justice						x
Leisure time activities (for respondents younger than 25)		x	x	x	x	x
Social integration	x	x	x	x	x	x
Social integration (special focus)			x		x	
Social Media						x
Health	x	x	x	x	x	x
Health (special focus)			x			x
Sports						x
Care	x	x	x	x	x	x
Pensions	x	x	x	x	x	x
Pensions (special focus)			x			

5 Structure of the scientific use file and its datasets

Benjamin Fuchs

The information collected in PASS is available as scientific use file (*SUF*). This chapter will give an introduction on how it is organised, the different types of datasets it includes on the individual and household level and the links between them.

Therefore, the first section of this chapter will deduce the *SUF*'s basic logic from the way households and its members are questioned in PASS. In doing so, it will be shown how the datasets of the *SUF* can be classified by their level (household or individual) and their type (register; cross section; weight or spell) and in which formats they are prepared (wide; long; spell).

Subsequently in the second section, we will focus on the datasets themselves. After a brief overview of the content of the *SUF*, the datasets will be presented in more detail, starting with the different types of datasets on the household level followed by the individual level.

5.1 Introduction to the scientific use file

5.1.1 Levels in the scientific use file

To understand the structure of the *SUF*, it is crucial to know that PASS collects information on the household as well as on the individual level and that these two levels are linked due to the survey design (see section 3). PASS surveys specific households and then questions the persons aged 15 and over living in these households at the time of the interview. The questioning of a household and its members starts by recording or updating the structure and other information concerning the whole household using the household questionnaire. After the household level information is collected, the household members suitable for individual interviews are known. PASS tries to question all persons up from the age of 15 with individual interviews. Because of this succession, where the household gives information about its members who are then targeted for individual interviews, each person in PASS is linked to a specific household in every single wave.

Due to the logic of the survey to collect information on the household level and on individuals living in these households, the *SUF* contains these levels as well. Therefore, each dataset of the *SUF* can be assigned to the household or the individual level.

5.1.2 Types of datasets in the scientific use file

The second criterion, by which the datasets of the *SUF* can be classified, is their type. The types of datasets that can be found on either level are attached to the contents of the survey (while the levels are attached to the surveys' basic logic). On each level, the *SUF* contains four different types of datasets (register; cross-sectional; weight; spell).

The *SUF* contains register datasets. The household register contains a list of all households that have ever been surveyed in PASS, while the person register contains a full list of all persons in these households. These register datasets provide basic information about the survey status of the household or person in every wave as well as additional wave-specific information.

While the register datasets contain only basic information about the household, their members and the respective survey status, the cross-sectional datasets of PASS contain most of the survey data collected during the interviews at the household and individual level, excluding the parts where the respondent was asked to report episodes (e. g. on the receipt of Unemployment Benefit II (UB II)). The cross-sectional data refers to the date of the interview it was collected – it represents the situation at a certain point in time.

PASS has a complex sample design which does not allow descriptive analyses without using weights. Therefore, the *SUF* contains weighting datasets on the household and the individual level. These datasets correspond to the cross-sectional datasets in their structure – they contain weights for each wave a household or person was surveyed in PASS that can be used to project the samples on the different populations (see section 12 on how to use the weights and section 8 on the weighting concept).

In addition, the *SUF* includes several spell datasets for information recorded in form of episodes. This way to collect data differs strongly from the cross-sectional concept described above. Therefore, it cannot be integrated directly in the cross-sectional datasets. When asked to report activities or events in form of episodes, the respondent had to fill a certain time period starting in the past and reporting all relevant activities or events up to the date of the interview. For each single episode the respondent had to report the begin date and end date and to give further information about its content. In each wave, several episodes can be recorded each of which refers to the period between its reported begin date and end date. Some of the periods may cover the time of interview and others may not. This kind of information is organised in spell datasets where each episode of the respondent forms a single observation.

5.1.3 Wide format, long format and spell format

As described above, the *SUF* contains four types of datasets (register; cross-sectional; weights; spells) on two levels (household; individual). These four types of datasets are prepared in three formats (wide format; long format; spell format).

The register datasets of PASS are prepared in wide format. This means that each unit is represented by exactly one observation in the dataset (= one row in the data matrix). Wave-specific information is allocated to these units in wave-specific variables. For waves, where no information is available for one unit the wave-specific variables are filled with specific missing value codes. Therefore, the observations of the register datasets uniquely present certain units and can be identified using a single key variable.

The cross-sectional and weighting datasets of PASS are both prepared in long format and not – as can be found in some other panel surveys – in separate annual files. Each wave a unit was surveyed, is represented by another observation in the dataset (= as many rows in the data matrix, as waves the unit was surveyed in). Thus, the wave-specific information can be found in wave-specific observations for the unit. Each variable, even if it is repeatedly collected in different panel waves, is only one column. Changes in the way a question is asked can lead to the decision that a new variable has to be integrated in the dataset. Variables surveyed only for certain waves are assigned the missing code "-9" for waves in which they were not surveyed. Therefore, the observations in the cross-sectional and the weighting datasets represent certain units in certain waves and can be identified using a combination of key variables for the unit and the wave.

The spell datasets of PASS are prepared in spell format. Each episode that was recorded for a unit, is represented by another observation in the dataset (= as many rows in the data matrix, as episodes reported by the unit). An episode can include information that was recorded in more than one wave when a current episode was updated in a following wave. Units that never reported an episode although they were successfully surveyed are not represented by an observation in the spell dataset. Units that reported more than one episode are represented by one observation per reported episode. Therefore, the observations in the spell datasets represent certain episodes of certain units and can be identified using a combination of key variables for the unit and the number of the spell.

5.2 Datasets of the scientific use file

The scientific use file of PASS consists of several datasets. As described above, these can be grouped by three criteria: level (household; individual), type (register; cross-sectional; weights; spells; factorial) and format (wide format; long format; spell format). Table 3 provides an overview of the datasets that are part of the *SUF* in wave 3 as well as their level, type and format. Each dataset will be described in more detail in the following sections, starting with the datasets on the household level followed by those on the individual level.

Table 3: Overview of the datasets of the scientific use file

Type	**Format**	**Name of dataset (information on waves and filenames in brackets) on ...**	
		Household level	**Individual level**
Register	wide	Household register *(hh_register)*	Person register *(p_register)*
Cross section	long	Household dataset *(HHENDDAT)*	Person dataset *(PENDDAT)* Children Dataset *(KINDER)*
		Household dataset on retirement provision (wave 3 only) *(HAVDAT)*	Person dataset on retirement provision (wave 3 only) *(PAVDAT)*
Weights	long	Household weights *(hweights)*	Person weights *(pweights)*
Spells	spell	Unemployment Benefit II spells *(alg2_spells)*	Biography spells (from wave 2) *(bio_spells)*
			1-euro-job spells (from wave 2) *(ee_spells)*
			Measure spells (wave 2 and 3 only) *(mn_spells)*
			Unemployment Benefit I spells (wave 1 only) *(alg1_spells)*
			Measure spells (wave 1 only) *(massnahmespells)*
Factorial	long		Vignette dataset (wave 5 only) *(VIGDAT)*

To describe the datasets in a layout that is easy to read, a standard table shown in Table 4 will be used. The meaning of the different categories was included in italic font and should be self-explanatory.

Table 4: Standard table for information on the characteristics of the dataset

Dataset:	Full name of the dataset , e. g. "Household register"
File name:	Filename of the dataset in the scientific use file, e. g. "*hh_register*"
Level:	Level of the dataset, e. g. "household"
Type:	Type of the dataset, e. g. "register"
Format:	Format of the dataset, e. g. "wide"
Data collected in waves:	Wave from which the dataset includes information, e. g. "1–6"
Integration of data from new waves:	Logic used to integrate information from new waves, e. g. "(1) Households that were surveyed for the first time are added as new observations. (2) New wave-specific variables are added. They include the information recorded in the lastwave."
Key variables:	All key variables included in the dataset, e. g. "(1) *hnr* (Household number) (2) *hnr$* (Household number in wave $)"
Pointer variables:	All pointer variables included in the dataset, e. g. "(1) *uhnr* (Original household number) (2) *pnrzp$* (Constant personal ID number of person who gave the household interview in wave $)"
One obs. (= row in data matrix):	What exactly is represented by one observation, e.g. "One household, that was at least once successfully surveyed in PASS."
One obs. (= row in data matrix) uniquely identified by:	Key variable, that uniquely identifies an observation, e. g. "*hnr*"
Topics:	Information on the topics covered by the dataset, e. g. "(1) Constant: sampling information (2) Wave-specific household information: households' survey status; size of household; number of synthetic benefit units; pointers"
Explanatory notes:	Notes that point out special characteristics or give additional information on the dataset, e. g. "Only households that were successfully surveyed at least once are included in the household register."

All datasets include key variables which are used to identify units and observations and to establish links to other datasets of the *SUF*. The key variables included in the dataset are listed in the corresponding tables (see "Key variables"). Further information about their meaning and on how to use them can be found in chapter 9. We strongly request the users of PASS to make themselves familiar with the structure of the datasets, their meaning and the key variables before combining different datasets.

A second group close to the key variables is the pointer variables. While the key variables are used to identify the same unit and link it between datasets, the pointer variables are used to establish links between different units, e. g. the variable "*uhnr*" (original household number) can be used to link a split-off household to its household of origin.

5.2.1 Household level datasets

Household register (hh_register)

Table 5: Characteristics of the household register dataset (*hh_register*)

Dataset:	Household register
File name:	*hh_register*
Level:	household
Type:	register
Format:	wide
Data collected in waves:	1–6
Integration of data from new waves:	(1) Households that were surveyed for the first time are added as new observations. (2) New wave-specific variables are added. They include the information recorded in the last wave.
Key variables:	(1) *hnr* (Household number) (2) *hnr$* (Household number in wave $)
Pointer variables:	(1) *uhnr* (Original household number) (2) *pnrzp$* (Constant personal ID number of person who gave the household interview in wave $)
One obs. (= row in data matrix):	One household that was at least once successfully surveyed in PASS.
One obs. (= row in data matrix) uniquely identified by:	*hnr*
Topics:	(1) Constant: sampling information (2) Wave-specific household information: households' survey status; size of household; number of synthetic benefit units; pointers
Explanatory notes:	Only households that were successfully surveyed at least once are included in the household register.

Household dataset (HHENDDAT)

Table 6: Characteristics of the household dataset (*HHENDDAT*)

Dataset:	Household dataset
File name:	*HHENDDAT*
Level:	household
Type:	cross section
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) Each newly recorded household interview is added as new observation in the dataset. (2) The newly recorded information is assigned to existing variables for this new observation. New variables are added if they were surveyed for the first time.
Key variables:	*hnr* (Household number) *welle* (Indicator for survey wave)
Pointer variables:	*uhnr* (Original household number)
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain household in a certain wave.
One obs. (= row in data matrix) uniquely identified by:	*hnr* + *welle*
Topics:	(1) Housing (2) General (2) Standard of living (3) Demography (4) Income (5) Child-care (6) Education and inclusion subsidies
Explanatory notes:	Only household interviews of households which were successfully surveyed according to the definition of PASS were included in the Dataset (see chapter 7.1 for definition). The dataset contains variables that are required to specify the “survey set” command in STATA (*psu, strata*). The dataset includes as many observations for a certain household, as the number of waves this household was successfully interviewed.

Table 7: Characteristics of the household dataset on retirement provision (*HAVDAT*)

Dataset:	Household dataset on retirement provision
File name:	*HAVDAT*
Level:	household
Type:	cross section
Format:	long
Data collected in waves:	wave 3 only
Integration of data from new waves:	In-depth information on retirement provisions was only collected in wave 3. Therefore, no data from new waves need to be integrated.
Key variables:	*hnr* (Household number) *welle* (Indicator for survey wave)
Pointer variables:	*uhnr* (Original household number)
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain household in wave 3.
One obs. (= row in data matrix) uniquely identified by:	*hnr* + *welle*
Topics:	(1) In-depth household information on retirement provisions.
Explanatory notes:	In-depth information on retirement provision was only collected in wave 3. The respective module of the household questionnaire was only asked for households where at least one person was 40 to 64 years old. The dataset contains observations for each household interviewed successfully in wave 3. In households for which no in-depth information on retirement provisions were collected, the survey variables were assigned the missing code “-3”.

Household weights (hweights)

Table 8: Characteristics of the household weights (*hweights*)

Dataset:	Household weights
File name:	*hweights*
Level:	household
Type:	cross section
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) Each wave a household is successfully interviewed is added as new observation in the dataset. (2) New weights are assigned to existing variables for this new observation.
Key variables:	*hnr* (Household number) *welle* (Indicator for survey wave)
Pointer variables:	–
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain household in a certain wave.
One obs. (= row in data matrix) uniquely identified by:	*hnr* + *welle*
Topics:	(1) Information on sample (2) Design weights for the total sample and the subsamples (3) Households participation probability in year of sampling (4) Projection factors for households of the total sample and the subsamples (5) Households reciprocal re-participation probability
Explanatory notes:	Only household interviews of households which were successfully surveyed according to the definition of PASS were included in the dataset (see chapter 7.1 for definition). The dataset includes as many observations for a certain household, as the number of waves this household was successfully interviewed.

Table 9: Characteristics of the Unemployment Benefit II spells (*alg2_spells*)

Dataset:	Unemployment Benefit II spells
File name:	*alg2_spells*
Level:	household
Type:	spells
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) New episodes that were reported in the last interview are added as new observations to the dataset. (2) Current spells from the time of the last interview were updated if the household has been interviewed. (3) The newly recorded information is assigned to existing variables. New variables are added if they were surveyed for the first time or if they refer to a certain wave (cross-sectional information as part of an UB II-episode).
Key variables:	*hnr* (Household number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	Episode during which a certain household received UB II.
One obs. (= row in data matrix) uniquely identified by:	*hnr + spellnr*
Topics:	(1) Information on UB II-recipiency (start date; end date; total amount of benefits per month; reason for beginning/end of recipiency) (2) Identification of household members receiving benefits (3) Benefit cuts (start date, end date, duration, reasons)
Explanatory notes:	Households that have never reported an episode UB II-recipiency are not represented by an observation in the dataset. The dataset includes as many observations for a certain household, as the number of episodes this household reported over the waves. An episode includes information that refers to the spell itself (e. g. the start date) as well as information that refers to a certain wave (e. g. the amount of benefits the household received in wave 3). These cross-sectional information are valid only for a certain point in time and can change while the episode continues. Therefore, the dataset contains cross-sectional variables referring to a certain wave. They are filled if the episode covers the respective wave and are otherwise assigned the missing code “-9”. The wave, a cross-sectional variable in the spells refers to, can be read from the variable labels.

5.2.2 Individual level datasets

Person register (p_register)

Table 10: Characteristics of the person register dataset (*p_register*)

Dataset:	Person register
File name:	*p_register*
Level:	individual
Type:	register
Format:	wide
Data collected in waves:	1–6
Integration of data from new waves:	(1) Persons, that are members of a surveyed household for the first time, are added as new observations. (2) New wave-specific variables are added. They include the information recorded in the last wave.
Key variables:	(1) *pnr* (Constant personal ID number) (2) *hnr$* (Household number in wave $) (3) *zplfd$* (Serial number of the target person in the household in wave $)
Pointer variables:	(1) *uhnr* (Original household number) (2) *zmhh$* (Constant personal ID number of target persons mother living in the same household in wave $) (3) *zvhh$* (Constant personal ID number of target persons father living in the same household in wave $) (4) *zparthh$* (Constant personal ID number of target persons partner living in the same household in wave $)
One obs. (= row in data matrix):	One person that was at least once a member of a successfully surveyed household in PASS
One obs. (= row in data matrix) uniquely identified by:	*pnr*
Topics:	(1) Constant: sampling information; information on persons sex and entry in the panel study (2) Wave-specific household information: household the person is a member of; serial number in the household (3) Wave-specific individual information: persons survey status; age; inclusion in the children dataset (4) Wave-specific synthetic benefit unit information: number, type and recipiency of the persons synthetic benefit unit (4) Wave-specific: pointers
Explanatory notes:	Only persons that were at least once members of a successfully surveyed household are included in the person register.

Person dataset (PENDDAT)

Table 11: Characteristics of the person dataset (*PENDDAT*)

Dataset:	Person dataset
File name:	*PENDDAT*
Level:	individual
Type:	cross-section
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) Each wave a person is successfully interviewed is added as new observation in the dataset. (2) The newly recorded information is assigned to existing variables for this new observation. New variables are added if they were surveyed for the first time.
Key variables:	*pnr* (Constant person ID number) *hnr* (Household number) *welle* (Indicator for survey wave)
Pointer variables:	*uhnr* (Original household number)
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain person in a certain wave.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *welle*
Topics:	(1) Demography (2) General (3) Social Media (4) Education (5) Attitudes and orientations (6) Perceptions about justice (7) Employment (8) Income (9) Contact to social security insitutions (10) 1-euro-jobs (11) Job-search (12) Standard of living (13) Social relations (14) Health (15) Sports (16) Care (17) Migration (18) Social origin
Explanatory notes:	The dataset includes as many observations for a certain person, as the number of waves this person was successfully interviewed.

Children dataset (KINDER)

Table 12: Characteristics of the children dataset (*KINDER*)

Dataset:	Children dataset
File name:	*KINDER*
Level:	individual
Type:	cross-section
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) Each wave for which at least one of the person-related questions of the HKI, HT or HBT modules was asked for the respective person is added as new observation in the dataset. (2) The newly recorded information is assigned to existing variables for this new observation. New variables are added if they were surveyed for the first time.
Key variables:	*pnr* (Constant person ID number) *hnr* (Household number) *welle* (Indicator for survey wave)
Pointer variables:	*zmhh* (Constant personal ID number of target persons mother living in the same household in the respective wave.) *zvhh* (Constant personal ID number of target persons father living in the same household in the respective wave.)
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain person in a certain wave.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *welle*
Topics:	(1) Demography (2) Child-care (3) Social participation (4) Education and inclusion subsidies
Explanatory notes:	This dataset is included in the SUF since wave 6. It contains variables with information about children living in the household which was stored in the household dataset until wave 5. The respective variables have been deleted in the household dataset. Only persons for whom one of the questions of the HKI, HT or HBT modules was asked within the household interview are contained in the children dataset. Consequently, it contains also persons above the age of 14. Due to the fact that no person weights are available for children under the age of 15, this dataset cannot completely and should not be merged with the person weights. This dataset in combination with the personal dataset does not yield representative results for children of different age groups. Examples for the usage of the children dataset are given in XXX.

Person dataset on retire provision (PAVDAT)

Table 13: Characteristics of the person dataset on retirement provision (*PAVDAT*)

Dataset:	Person dataset on retirement provision
File name:	*PAVDAT*
Level:	individual
Type:	cross-section
Format:	long
Data collected in waves:	wave 3 only
Integration of data from new waves:	In-depth information on retirement provisions was only collected in wave 3. Therefore, no data from new waves need to be integrated.
Key variables:	*pnr* (Constant personal ID number) *welle* (Indicator for survey wave)
Pointer variables:	–
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain person in wave 3.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *welle*
Topics:	(1) In-depth individual information on retirement provisions.
Explanatory notes:	In-depth information on retirement provision was only collected in wave 3. The respective module of the persons' questionnaire was only asked for persons who were 40 to 64 years old or had a partner of this age. The dataset contains observations for each person interviewed successfully in wave 3. For persons for whom no in-depth information on retirement provisions were collected, the survey variables were assigned the missing code "-3".

Person weights (pweights)

Table 14: Characteristics of the person weights (*pweights*)

Dataset:	Person weights
File name:	*pweights*
Level:	individual
Type:	cross-section
Format:	long
Data collected in waves:	1–6
Integration of data from new waves:	(1) Each wave a person is successfully interviewed is added as new observation in the dataset. (2) New weights are assigned to existing variables for this new observation.
Key variables:	*pnr* (Constant personal ID number) *welle* (Indicator for survey wave)
Pointer variables:	–
One obs. (= row in data matrix):	Cross-sectional information, regarding a certain person in a certain wave.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *welle*
Topics:	(1) Information on sample (2) Projection factors for persons of the total sample and the subsamples (3) Persons reciprocal re-participation probability
Explanatory notes:	The dataset includes as many observations for a certain person, as the number of waves this person was successfully interviewed.

Biography spells (bio_spells)

Table 15: Characteristics of the Biography spells (*bio_spells*)

Dataset:	Biography Spells
File name:	*bio_spells*
Level:	individual
Type:	spells
Format:	spell
Data collected in waves:	2–6 (integration since wave 4)
Integration of data from new waves:	(1) New episodes that were reported in the last interview are added as new observations to the dataset. (2) Current spells from the time of the last interview were updated if the person has been interviewed. (3) The newly recorded information is assigned to existing variables. New Variables are added if they were surveyed for the first time or if they refer to a certain wave (cross-sectional information as part of a biographical episode).
Key variables:	*pnr* (Constant personal ID number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	One biographical episode.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *spellnr*
Topics:	(1) Information on employment with an income of more than 400 euros (start date; end date; occupational status; number of staff/cultivated area; supervisor function; number of supervised employees, temporary/fixed-term contract and conversion; working hours; temporary work; public sector; no. of employees in establishment/local office; reason for termination of contract; way of taking notice of the job offer; income; various measures of occupational status and prestige; sector (2) Information on registered unemployment/employment or training measure participation (start date, end date, reason for exit) and recipiency of UB I during an episode of registered unemployment (start date; end date; total amount of benefits per month) (3) Information on other biographical episodes (unregistered unemployment; educational training; military/alternative service; housewife/househusband; maternity protection/parental leave; retirement; sickness/job-related incapacity; self-employment; other activities)
Explanatory notes:	Employment with an income of more than 400 euros was recorded as part of the persons' questionnaires' biography module. Since wave 4, in addition to episodes of employment, the respondent was asked for episodes of registered unemployment and other biographical episodes as mentioned above. This is different to the procedure in wave 2-3, where employment spells, unemployment spells, and other biographical episodes were questioned seperately. Also, they were stored seperately in the datasets *et_spells*, *al_spells* and *lu_spells*. These datasets were integrated into the *bio_spells* dataset in wave four and are not contained in the SUF anymore. Persons who have never reported an episode of employment, unemployment or other biographical episodes are not represented by an observation in the dataset. The dataset includes as many observations for a certain person, as the number of episodes this person reported over the waves. An episode includes information that refers to the spell itself (e. g. the start date) as well as information that refers to a certain wave (e. g. the simple classification of the occupational status in wave 6). These cross-sectional information are valid only for a certain point in time and can change while the episode continues. Therefore, the dataset contains cross-sectional variables referring to a certain wave. They are filled if the episode covers the respective wave and are otherwise assigned the missing code "-9". The wave, a cross-sectional variable in the spells refers to, can be read from the variable labels.

Measure spells (from wave 2) (mn_spells)

Table 16: Characteristics of the measure spells (*mn_spells*)

Dataset:	Measure Spells (from wave 2)
File name:	*mn_spells*
Level:	individual
Type:	spells
Format:	spell
Data collected in waves:	2–3
Integration of data from new waves:	(1) New episodes that were reported in the last interview are added as new observations to the dataset. (2) Current spells from the time of the last interview were not updated. (3) The newly recorded information is assigned to existing variables.
Key variables:	*pnr* (Constant personal ID number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	Episode during which a certain person participated in a certain employment/training measure.
One obs. (= row in data matrix) uniquely identified by:	*pnr + spellnr*
Topics:	(1) Information on measure (start date; end date/duration for completed and current measures; type of measure; subject of measure; reason for ending measure prematurely; initiative for participation; assessment of measure; hours per week; requirements; identical work as permanent employees; social education worker present; sector)
Explanatory notes:	In wave 2, the concept for surveying participation in employment and training measures was reworked because in the concept of wave 1 it proved difficult to identify clearly the exact type of the measure (with the exception of the one-Euro-jobs, which were recorded directly). In wave 2 the type of measure in which a person had participated was first recorded directly using multiple choice questions. Then further information was collected in the form of looped sequences of questions about the reported measure types. As a special characteristic, different types of end dates/durations were asked for the measure episodes. For measures, that were already completed, the real end date/duration was recorded. For current measures, in which the respondent still participated, the intended end date/duration was recorded. The later were marked as right-censored using the variable "*zensiert*". In contrast to the employment, unemployment and UB II spells, the current measure spells were not updated in the following interview. Instead, spells that had not been completed at the time of the interview stay right-censored. Therefore, the meaning of a right-censored spell differs from other spell datasets. Here, a spell that is right-censored does not mean that it is current at the time of the respondents' last interview but that it was current at the time of the interview when it was reported. The wave a measure spell was reported in can be identified using the wave indicator "*spwelle*" included in the dataset. Therefore, a right-censored measure spell was current in the wave indicated by "*spwelle*". Persons who have never reported an episode of measure participation are not represented by an observation in the dataset. The dataset includes as many observations for a certain person as the number of episodes this person reported over the waves. This dataset has not been continued since wave 4, ever since information about measure participation is stored in the *ee_spells* dataset.

Unemployment Benefit I spells (alg1_spells)

Table 17: Characteristics of the Unemployment Benefit I Spells (wave 1 only) (*alg1_spells*)

Dataset:	Unemployment Benefit I Spells (wave 1 only)
File name:	*alg1_spells*
Level:	individual
Type:	spells
Format:	spell
Data collected in waves:	1
Integration of data from new waves:	(1) Episodes of UB I recipiency were only recorded directly in wave 1. Therefore, no data from new waves need to be integrated.
Key variables:	*pnr* (Constant personal ID number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	Episode during which a certain person received UB I.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *spellnr*
Topics:	(1) Information on UB I-recipiency (start date; end date; total amount of benefits per month)
Explanatory notes:	Episodes of UB I-recipiency were only recorded directly in wave 1. Starting with wave 2, the information on times when the respondent received this benefit was recorded as part of the episodes of registered unemployment. From wave 2 on, information on UB I-recipiency can be found in the unemployment spell dataset. Persons who have not reported an episode of UB I-recipiency in wave 1 are not represented by an observation in the dataset. The dataset includes as many observations for a certain person, as the number of episodes this person reported in wave 1.

Measure Spells (massnahmespells)

Table 18: Characteristics of the measure spells (wave 1 only) (*massnahmespells*)

Dataset:	Measure spells (wave 1 only)
File name:	*massnahmespells*
Level:	individual
Type:	spells
Format:	spell
Data collected in waves:	1
Integration of data from new waves:	The concept of wave 1 to survey measure participation was reworked in wave 2. Therefore, no data from new waves need to be integrated.
Key variables:	*pnr* (Constant personal ID number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	Episode during which a certain person participated in a certain employment/-training measure.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *spellnr*
Topics:	(1) Information on measure (start date; duration for completed and current measures; type of measure; reason for ending measure prematurely; initiative for participation; assessment of measure; hours per week; requirements; identical work as permanent employees; social education worker present)
Explanatory notes:	In wave 2, the concept for surveying participation in employment and training measures was reworked because in the concept of wave 1 it proved difficult to identify clearly the exact type of the measure (with the exception of the one-Euro-jobs, which were recorded directly). Because of the extent of the changes, the information recorded from wave 2 on could not be integrated in the measure spell dataset of wave 1. Persons who have not reported an episode of measure participation in wave 1 are not represented by an observation in the dataset. The dataset includes as many observations for a certain person, as the number of episodes this person reported in wave 1.

1-euro-job spells (ee_spells)

Table 19: Characteristics of the 1-euro-job spells (*ee_spells*)

Dataset:	1-euro-job spells
File name:	*ee_spells*
Level:	individual
Type:	spells
Format:	spell
Data collected in waves:	2–6
Integration of data from new waves:	(1) New episodes that were reported in the last interview are added as new observations to the dataset. (2) Current spells from the time of the last interview were updated if the person has been interviewed. (3) The newly recorded information is assigned to existing variables. New variables are added if they were surveyed for the first time or if they refer to a certain wave (cross-sectional information as part of an 1-euro-job episode).
Key variables:	*pnr* (Constant personal ID number) *spellnr* (Spell number)
Pointer variables:	–
One obs. (= row in data matrix):	Episode during which a certain person participated in or got an offer for an 1-euro-job.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *spellnr*
Topics:	(1) Information on the 1-euro-job (date of the offer; start date and end date/duration for finished and current 1-euro jobs; (non-)participation; reasons for non-participation; UB II cuts due to non-participation; premature ending and reasons for it; hours per week; identical work as permanent employees; requirements; subjective assessments of the 1-euro job
Explanatory notes:	In wave 4, the concept for surveying participation in employment and training measures was reworked and refers exclusively to 1-euro-jobs since then. Therefore, the *mn_spells* dataset is not continued. The starting point in wave 4 was January 2009 for all respondents. 1-euro-job episodes from the datasets *massnahmespells_spells* or *mn_spells* have not been integrated into the *ee_spells* dataset. Persons who have never reported an offer or an episode of an 1-euro-job are not represented by an observation in the dataset. The dataset includes as many observations for a certain person as the number of offers or episodes this person reported over the waves. Consequently, there are also spells contained if the person did not participate in an (offered) 1-euro-job.

Vignette dataset (VIGDAT)

Table 20: Characteristics of the vignette dataset (*VIGDAT*)

Dataset:	Vignette dataset
File name:	*VIGDAT*
Level:	individual
Type:	factorial
Format:	long
Data collected in waves:	5
Integration of data from new waves:	The factorial survey module was only questioned in wave 5. Therefore, no data from new waves need to be integrated.
Key variables:	*pnr* (Constant personal ID number) *vignr* (vignette number)
Pointer variables:	-
One obs. (= row in data matrix):	Factorial survey information, regarding a certain person with a certain vignette.
One obs. (= row in data matrix) uniquely identified by:	*pnr* + *vignr*
Topics:	(1) Characteristics of the job offer (vignette dimensions) (2) Assessment of the job offer
Explanatory notes:	

6 Variable types and their names

Arne Bethmann

6.1 General issues

For naming the variables of the dataset we considered two main alternatives from which we had to choose one. The first option is naming the variables in accordance with their respective order in the questionnaire, as is done in the German Socio-Economic Panel (GSOEP), for example. The advantage of this type of naming convention is that the items corresponding to the variables are easy to find in the questionnaire, which significantly enhances the value of the questionnaire as a documentation instrument. The central disadvantage of this approach is that identical items are given different names due to changes in the order of questions in the questionnaire, resulting in considerable preparation being required for compiling and, if necessary, renaming the required variables even for simple trend analyses, as more and more panel waves become available.

The second main alternative is allocating independent variable names, which are kept constant across waves (apart from a wave indicator if necessary). The advantages and disadvantages of this strategy are opposite to those of the first alternative: identifying the variables corresponding to an item across waves is unproblematic, whereas using the questionnaire as a documentation instrument becomes more difficult, as it is no longer possible to derive the position of an item in the questionnaire from the variable name. In our opinion, the advantages of fixed variable names clearly outweigh the disadvantages in a long-term panel study. Moreover, the decision in favour of organising the data in long format as described above requires the use of uniform variable names.

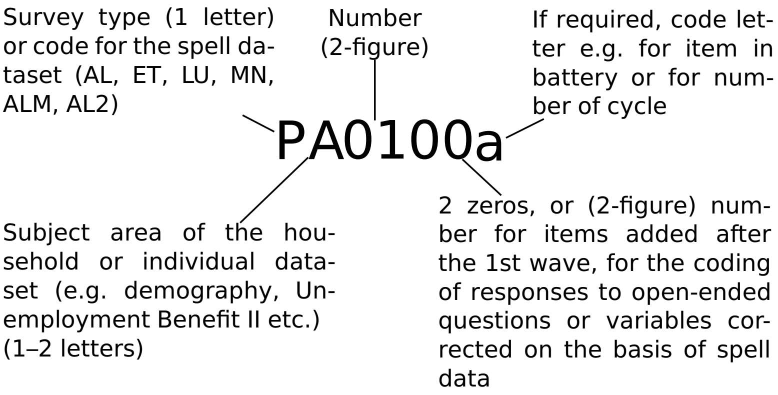
6.2 Variable types

The codebook distinguishes between three different types of variables:

6.2.1 System variables

System variables are variables created in the course of the survey process. They can be used, firstly, to comprehend the filters documented in the questionnaire. At least some of the system variables can also be of interest from a content-related or methodological point of view, for example the interview mode or the number of children of a certain age group living in the household. System variables are allocated individual names, for which lower-case letters and numbers are combined in some cases. The system variables also include the weights.

Figure 1: The variable naming scheme


Survey type (1 letter) or code for the spell dataset (AL, ET, LU, MN, ALM, AL2)
Number (2-figure)
If required, code letter e.g. for item in battery or for number of cycle
PA0100a
Subject area of the household or individual dataset (e.g. demography, Unemployment Benefit II etc.) (1–2 letters)
2 zeros, or (2-figure) number for items added after the 1st wave, for the coding of responses to open-ended questions or variables corrected on the basis of spell data

6.2.2 Surveyed variables

Surveyed variables are variables that were collected in this form directly in the questionnaire. These variables are given entirely new, abstract variable names. The concept behind this naming process is illustrated in Figure 1 using an example.

- The 1st letter of the variable name indicates the questionnaire level, i. e. household or individual dataset, by means of the letter H or P (upper-case) respectively.
- This is followed by one or two upper-case letters which indicate the subject area to which the variable belongs (see Tables 21 and 22 for a complete list).
- In the datasets which are processed in spell form, there is no introductory P or H. Instead, the variables in these datasets are given a uniform subject-based name consisting of two or three letters or two letters and one number.
- The introductory letter combination is then followed by two consecutively allocated numbers, which indicate the number of the question within the subject area.
- These two numbers are followed by two zeros, which are intended to permit the addition of further variables in later waves. Also this option has been used in cases where a second variant including coded information from an open-ended survey question or response category has been made available in addition to the original version of the variable. The final zero is changed to “1” for these variables (e. g. *PA0101a* instead of *PA0100a*).
- In the case of variables for items from multi-item batteries or in a looped sequence of questions, a further lower-case letter may be added to identify the item or the current cycle within the loop.

6.2.3 Generated variables

The generated variables in the strict sense are aggregated from various other variables, e. g. from open-ended and categorical income measures, or they are even more complex

Table 21: List of subject-related indicators used in the variable names (cross sections)

Individual level		Household level	
Code	**Subject area**	**Code**	**Subject area**
PD	Demography	HW	Housing
PA	General	HA	General
PSM	Social Media	HLS	Standard of living
PB	Education	HD	Demography
PEO	Attitudes and orientations	HEK	Income
PGR, PPG	Perceptions about justice	HKI	Child-care
PET	Employment	HT	Social participation
PEK	Income	HBT	Education and inclusion subsidies
PTK	Contact to social security institutions		
PEE	1-euro-jobs		
PAS	Job-search		
PLS	Standard of living		
PSK	Social relations		
PG	Health		
PSB	Sports		
PP	Care		
PMI	Migration		
PSH	Social origin		

constructs such as equivalised household income or classifications for education (such as ISCED or Casmin) or status (e. g. EGP, ESEC). Generated variables in this strict sense are allocated individual names that are as clear and memorable as possible, in lower-case letters. For an overview of the generated variables, see chapter 13.

Another group of generated variables includes those in which information from open-ended survey questions or response categories was added to another (closed) variable. Although these variables are, strictly speaking, also generated variables and are classified as such in the frequency tables of the codebook, they are not given clear names. Instead their names are based on those of the original variable, but with a “1” as the final number rather than a “0”.

Table 22: List of subject-related indicators used in the variable names (spells)

Individual level		Household level	
Code	Subject area	Code	Subject area
ET	Employment with earnings of more than 400 € per month since January 2005 (spell data, individual-level data, from wave 2 onwards)	AL2	Receipt of Unemployment Benefit II (spell data, household level)
AL	Spells of registered unemployment and receipt of Unemployment Benefit I since January 2005 (spell data, individual level, from 2nd wave onwards)		
LU	Other activities since January 2005 (spell data, individual-level data, from wave 2 onwards)		
EE	1-euro-jobs (spell data, individual level, from wave 4 onwards)		
MN	Employment and training measures (spell data, individual level, waves 2 and 3 only)		
ALM	Employment and training measures (spell data, individual level, wave 1 only)		
AL (wave 1)	Receipt of Unemployment Benefit I (spell data, individual level, wave 1 only)		

7 Data editing

Daniel Gebhardt

The Scientific Use File (*SUF*) of PASS is the product of an intensive data editing process. In its course, the raw data collected by the field institute in a certain wave is checked, answers to open-ended survey questions are coded, variables are generated and the data is integrated into the datasets of the *SUF*. Although this process is improved and adjusted for each wave, its basic logic and the succession of its steps stay the same over time. While the wave-specific procedures are described in the data reports (see for example Berg et al. 2013c for the data editing of wave 6) this section will focus on giving an overview of the important steps and their succession.

The data editing of the first two waves was performed at the Institute for Employment Research (IAB). With wave 3, the Institut für Angewandte Sozialwissenschaft (infas), the new field institute of PASS took over this task.[5] To ensure that this change in who edits the data would not result in a change in procedures and inconsistency in the datasets of the *SUF*, several precautions were taken. First, the new contract with infas stated as a condition that all steps of the data editing process had to be carried out in the same order and in an analogue way as in the previous waves. infas was therefore provided with the relevant syntax files and datasets of wave 2 as well as with a documentation of each step. Second, the process of data editing was accompanied by continuous coordination between infas and the IAB. Important decisions e. g. on problematic household structures or on the integration of spell datasets were made after consulting the IAB. In addition, the IAB was open for discussion and requests during the whole process. Third, after the *SUF* of wave 3 was finished, the final datasets were subject to a final check by the IAB regarding their structure and content.

Besides this, the logic and succession of the data editing process stayed the same over the waves. It can be divided into the following steps:

[5] The contract with the former field institute TNS Infratest was initially limited for three waves. As a consequence, the field work from wave 4 on had to be put out on a request for proposals in which the IAB decided to include the data editing starting with wave 3. Therefore, infas as the new field institute of PASS from wave 4 on also carried out the data editing of wave 3.

Table 23: Overview of the steps involved in editing the data of PASS

No.	Step of the procedure
1	Check of the household structure of re-interviewed households
2	Removal of problematic/incomplete interviews (household and/or individual level)
3	Integration of individual dataset and senior citizens' dataset
4	Correction of the household structure of re-interviewed households
5	Filter checks at the household level
6	Construction of a household grid dataset and plausibility checks
7	Generation of the synthetic benefit units (see description of variables in wave-specific data reports)
8	Generation of new control variables on the basis of the household data following filter checks and the household grid dataset after plausibility checks
9	Filter checks at the individual level
10	Coding of information from open-ended survey questions
11	Plausibility checks of the household and individual-level data (excluding spell data)
12	Preparation, plausibility checks and construction of the spell datasets
13	Simple variable generations
14	Complex variable generations
15	Generation of the data structure for the scientific use file (household dataset, individual dataset, register dataset)
16	Anonymisation
17	Final check of the *SUF* datasets

7.1 Structure checks

First, the household structure of re-interviewed households was compared to the structure reported in the previous interview in order to identify and, if necessary, correct implausible or problematic changes in the household composition and errors in the allocation of the individual interviews to their respective position in the household. For observing the households in the longitudinal section it is essential that the individuals are assigned consistently to their position in the household and that the respondents can be identified clearly across the waves. The same personal identification number must not be allocated to different individuals in different waves.

If the correct household composition was unclear, all of the interviews conducted with the household in this wave were removed from the *SUF*. If one of the individual interviews was conducted with the wrong person but without any further problems emerging in the household composition, then just the individual interview was removed.

The wave-specific data reports give an overview of the checks carried out to identify problematic cases (e. g. see Berg et al. 2013c for wave 6). The net variables in the household register *(hnettok*, hnettod*)* and person register datasets *(pnettok*, pnettod*)* provide information about removed interviews over the waves. Please note that not all deleted interviews can be identified in the *SUF* due to the logic of the register files.[6]

[6] In PASS, the register files of the *SUF* are net files. Therefore, the household register contains all households that have ever been successfully surveyed. The person register contains all persons living in the households at the time of the interview. Removed interviews from households or persons that are not included in the

Second, incomplete interviews at the household and individual level were not included in the *SUF*[7] as well as interviews from households which were regarded as not successfully surveyed according to the definition of PASS (see Table 24).[8] These cases were not documented in the register datasets because they were not regarded eligible in the first place, in contrast to the removed interviews described above.

Table 24: Interviews, at least required for a household to be regarded as successfully surveyed in PASS

Type of household	Household level interwiew	Individual level interwiew(s)
new household (household was interviewed for the first time and drawn for the initial sample or a refreshment sample)	yes (completed)	yes (at least one completed)
re-interviewed household (household was already interviewed in a previous wave of PASS)	yes (completed)	none required
new split-off household (household was interviewed for the first time and is a split-off from another household in PASS)	yes (completed)	none required

7.2 Filter checks and assignment of standardised codes

Every surveyed variable in the *SUF* datasets was filter-checked. During these checks filter errors were marked and standardised missing codes were assigned. Table 25 gives an overview of the standardised codes used in PASS:

Table 25: Overview of standardised codes used in PASS

Code	Explanation
-1	"Don't know"
-2	"Details refused"
-3	"Not applicable (filter)" (question not asked due to filter)
-4	"Question mistakenly not asked" (question should have been asked)
-5	Question-specific code No. 1, only allocated as required
-6	Question-specific code No. 2, only allocated as required
-7	Question-specific code No. 3, only allocated as required
-8	"Implausible value"
-9	"Item not administered in wave"
-10	"Item not administered in questionnaire version"

register datasets cannot be traced in the *SUF* (e. g. removed first-time interviews of households from the refreshment sample or individual interviews in these households).

[7] Therefore, the datasets of the *SUF* do not include interviews that were canceled before the respondent finished the questionnaire.

[8] Because the definition of "successfully surveyed" differs between the types of households, the *SUF* contains households without interviews at the individual level in certain waves.

The standardised codes shown above can be divided into the following groups:

- Missing values due to direct answers of the respondent (“-1”, “-2”)
- Missing values due to filters (or problems with filters) (“-3”, “-4”)
- Question-specific codes (“-5”, “-6”, “-7”)
- Missing values due to implausible answers of the respondent (“-8”)
- Missing values due to questions not included in the questionnaire/wave (“-9”, “-10”)

With the exception of implausible answers which were identified later (see later on in this section), the other groups were treated during this step of data editing. The correct operation of the filters was checked and the system missings were replaced. Therefore, the variables of the raw datasets were examined step by step in the order in which they were recorded. Hereby, the codes “-3” and “-4” were assigned. A variable was set to “-3” (not applicable) if the question had not to be asked due to a filter condition. Questions that were asked even though they should not have been, were corrected to “-3” too.[9] While in this case falsely recorded information could be corrected (that is set to “-3”) easily, information could not be added to correct missing answers. If an item was not surveyed although it should have been according to the relevant filter condition, the missing code “-4” (question mistakenly not asked) was allocated to mark these cases. The codes “-1” and “-2” were assigned as standard values for “Don’t know” and “Details refused” recorded during the interview. The codes “-5” to “-7” are question-specific codes. These can either be specific missing codes (e. g. “Not applicable, not available for the labour market”), or special categories for valid values (e. g. a category for an income above €99,999 in the open question on income). These codes were only allocated as required. The missing codes for items that were not included in a specific questionnaire or wave were allocated. The code “-9” was assigned, if a certain item was not surveyed in a specific wave. Due to the dataset being prepared in long format (see section 5.1.3), variables that were not surveyed in a specific wave were given the value “-9” for the observations in that wave.[10] The code “-10” can be used to account for differences between the questionnaire versions, in other words between the standard questionnaire and the senior citizens’ questionnaire or between the two versions of the household questionnaire (only from wave 1 to 3).

[9] That is, they were set to the value they would have received if there had not been a problem with the filter condition (e. g. detailed information on vocational training should only be recorded, if the respondent stated that he/she has a vocational qualification. If it was recorded anyway, the variable was set to “-3”). In this case falsely recorded information was replaced by “-3”.

[10] For example variables that were surveyed for the first time in the 2nd wave were retroactively coded “-9” for observations of wave 1. On the other hand, variables only surveyed in the 1st wave were set to “-9” for the observations of the following waves.

8 Weighting

Mark Trappmann

This chapter contains information on the concept and process of constructing and calculating the weights. Information on how to use the weights can be found in section 12.

8.1 Initial weights

PASS consists of multiple subsamples (compare section 3.2): An initial recipient sample, a population sample, a refreshment sample for the recipient sample in each wave from wave 2 and two replenishment samples introduced in wave 5. The weighting process for each sample in the wave that the sample was first included consists of three stages:

- In the first stage, design weights are produced for the gross sample.
- Subsequently, non-response is modelled in the second stage.
- Finally, in the third stage the weights are calibrated.

8.1.1 Stage 1: design weighting

The design weights are reciprocal selection probabilities for the gross sample. Which procedure has been used to generate the weights is described in detail in Rudolph/Trappmann (2007) and in section 6.1 of Berg et al. (2013b) for the wave 5 replenishment sample of the general population. The design weights are contained in the dataset *hweights*. The individual design weights supplied are:

dw_ba Design weight of a household in the BA sample (population: households in which there was at least one benefit unit in joint receipt of benefits in accordance with Social Code Book II in any July since 2006)

dw_mi Design weight of a household in the general population sample (population: households in the Federal Republic of Germany)

dw Design weight of a household in the total sample (population: households in the Federal Republic of Germany)

8.1.2 Stage 2: modelling of nonresponse

With the aid of two logit models, the participation probability is estimated for all households in the gross sample. The first logit model explains the probability of a contact. The second logit model explains the participation (at least the household interview and one complete personal interview) conditional on a successful contact. These logit models are estimated separately for each subsample. The set of variables used in these propensity models has changed over the first five waves. In the wave 1 general population sample micro geographical variables supplied by Microm as well as regional indicators like state and municipal size were used. For details the user is referred to the wave specific reports by Hartmann et al. (2008), Büngeler et al. (2009) and Berg et al. (2011, 2013a,b,c). In the case of the models for the BA samples, additional characteristics on the level of the benefit unit from the sampling frames (A2LL or XSozial) could be used. From wave 3 on only variables on Microm variables were dropped and only variables on the level of the benfit unit, regional variables and the number of contact attempts were used for these samples. For the wave 5 population replenishment sample age, gender and nationality were used in addition to regional predictors.

The dataset *hweights* contains the variable *prop_t0*. This is the product of the predicted probabilities of the two models.

Dividing the design weights by the estimated participation probabilities yields the modified design weights, which formed the starting point for the third stage – calibration.

8.1.3 Stage 3: calibration

A detailed documentation of the calibration process of waves 1 and 2 can be found in Kiesl (2010). The calibration procedures and results reported by TNS Infratest in the method and field reports (Hartmann et al. 2008, Büngeler et al. 2009) are not the ones used for the weights in the scientific use file. The calibration of waves 3 to 6 is detailed in the data reports by infas (Berg et al., 2011, 2013a,b,c). We therefore merely outline the basic procedure here. This section will deal with the calibration of the initial samples in wave 1 only, because at later waves refreshment samples are not calibrated separately, but only within the calibration of the complete samples (see section 8.2.7 and 8.2.8).

Household level (wave 1) In an initial step, the two subsamples and the total sample were calibrated to official statistics at the household level.

The total and BA weights for benefit recipients in the two samples were calibrated to benchmark statistics from the Federal Employment Agency (reporting month July 2006). The total and Microm weights were additionally calibrated to benchmark statistics on private households in Germany for 2007 from the Federal Statistical Office. The benchmark figures used are detailed in Kiesl (2010).

All weights are household weights. The BA statistics, however, are based on values at the level of benefit units. The link is created using the synthetic benefit units, generated as described in the data report for wave 1 (Christoph et al. 2008: 49, variable *bgnr1* in the *p_register* dataset). Households are initially broken down into synthetic benefit units. The characteristics used for the calibration process are then generated at the benefit unit level. This also includes the characteristic of whether the benefit unit was receiving Unemployment Benefit II as of the sampling date. After calibration, multiplying the characteristics of all benefit units in receipt of benefits as of the sampling date by the projection factors for households yields the benchmark figures. Separate benefit units in receipt of benefits within one household are therefore always given the same projection factors.

It is not always possible to determine accurately the benefit receipt of a household or even of a benefit unit. As much data as possible is therefore provided in order to enable users to make independent decisions. Thus, for instance, the variable *alg2samp* at the household level is supplied in the *hh_register* dataset. This variable contains the benefit receipt as of the sampling date for all households in the categories: *0 no receipt, 1 receipt, 2 no receipt according to survey (but included in BA sample and thus receipt according to register data), 3 receipt unclear from survey (but included in BA sample and thus receipt according to register data), 4 receipt unclear from survey (general population sample)*. In addition, every user can generate this variable him/herself using the unemployment benefit II spell data (*alg2_spells* dataset). Other useful variables are *AL20600* and *AL20700a–o* (for which members does the household receive benefits?).

To generate the weights, however, a clear decision is needed on which benefit units should be regarded as being in receipt of Unemployment Benefit II on the sampling date. The decisions upon which the weighting is based can be explained as follows:
At the household level it was decided that:

1. All households from the BA sample (*sample=1*) were in receipt of benefits as of the sampling date even if they denied this, provided at least one person aged between 15 and 64 lives in the household.
2. Households from the general population sample for which benefit receipt can not be clearly established on the basis of the survey data are regarded as households receiving Unemployment Benefit II for the purpose of weighting if they report ever having received Unemployment Benefit II (*HA0300=1*) and if the start or end date of at least one observation lies in 2006 (in cases of an undetermined end or start).

Transferring from the household to the benefit unit level is wrought with even greater uncertainty. The reason for this is that it is not possible to obtain reliable retrospective information on which parts of the household received benefits in July 2006. In most cases, the entire household consists of only one benefit unit, making the question redundant, as the benefit unit receives benefits precisely when the household does so. In cases where the household consists of more than one benefit unit, the following approach was selected:
The information as to which individuals the household is currently receiving benefits for (*AL20600* and *AL20700a–o*) was used. A benefit unit was regarded as receiving benefits if at least one of its members was reported as a benefit recipient. In a household with more than one benefit unit and with no information as to which individuals the household is receiving benefits for (e. g. because the questionnaire responses state that no benefits are being claimed), all of the synthetic benefit units were regarded as being in receipt of benefits. The result of this generation is contained in the variable *bgbezs1* in the *p_register* dataset.

The weights following calibration at the household level are also contained in the *hweights* dataset.

wqbahh	calibrated household weight of the BA sample
wqmihh	calibrated household weight of the general population sample
wqhh	calibrated household weight of the total sample

Person level Following the calibration at the household level, the individuals who gave a personal or senior citizen's interview were calibrated to benchmark statistics at the individual level. The calibrated household weights were the starting point for this step.

The total and BA weights for benefit recipients in both subsamples were calibrated to benchmark statistics from the Federal Employment Agency (reporting month July 2006). The total and general population sample weights were additionally calibrated to benchmark statistics from the Federal Statistical Office on private households in Germany for 2007. The benchmark figures used are detailed in Kiesl (2010).

Senior citizens' interviews were calibrated to population statistics in the same way as the standard personal interviews. The BA statistics, however, do not contain figures on the number of senior citizens in households receiving benefits. Nor do they identify individuals living in households receiving benefits who are not part of a benefit unit. It was therefore impossible to obtain the BA person weights for these individuals by means of calibration. The participation probability of these individuals, given that their household takes part in the survey, was estimated using a logit model with the following covariates: number of individuals aged 15 and over in the household; interview mode; age and gender. The modified design weight was subsequently divided by this value.

The calibrated person weights are contained in the *pweights* dataset.

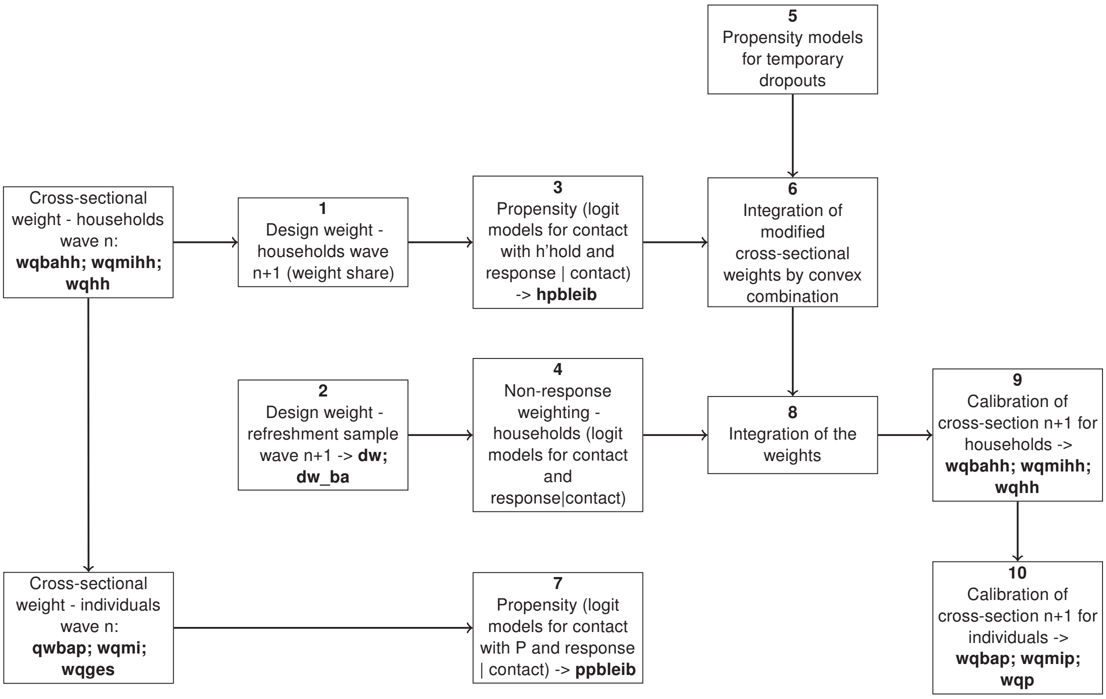
wqbap	calibrated person weight of the BA sample
wqmip	calibrated person weight of the Microm sample
wqgesp	calibrated person weight of the total sample

8.2 Construction of the weights from wave 2 onwards

The starting points for the weighting procedure for the second wave and for the longitudinal section from wave 1 to wave 2 are the cross-sectional weights from wave 1 for households and individuals. More generally, the starting points for the weighting procedure for the (n+1)th wave and for the longitudinal section from wave n to wave n+1 are the cross-sectional weights from wave n for households and individuals. In wave n ($n \geq 1$) each household had two weights, *wqhh* (calibrated total weight) and – depending on the sample – *wqbahh* (calibrated BA weight) or *wqmihh* (calibrated general population sample weight), and each individual also had two weights, *wqp* and – depending on the sample – *wqbap* (calibrated BA weight) or *wqmip* (calibrated general population sample weight). All four weights are updated for the following wave (wave n+1). Figure 2 shows the steps of the weighting procedure, which are explained below. This section is meant as a comprehensive overview. For waves 3–6 chapter 6 of each wave specific data report (Berg et al., 2011, 2013a,b,c) contains details about the exact models and variables used and on model coefficients.

For details on wave 2 the reader is referred to Büngeler et al. (2009). This section does not include a description of the integration of weights for replenishment samples with the ongoing panel samples. Section 8.3 is devoted to this special case.

Figure 2: Generation of the weights for wave n+1 given the weights of wave n



5 Propensity models for temporary dropouts
Cross-sectional weight - households wave n: wqbahh; wqmihh; wqhh
1 Design weight - households wave n+1 (weight share)
3 Propensity (logit models for contact with h'hold and response | contact) -> hpbleib
6 Integration of modified cross-sectional weights by convex combination
2 Design weight - refreshment sample wave n+1 -> dw; dw_ba
4 Non-response weighting - households (logit models for contact and response|contact)
8 Integration of the weights
9 Calibration of cross-section n+1 for households -> wqbahh; wqmihh; wqhh
Cross-sectional weight - individuals wave n: qwbap; wqmi; wqges
7 Propensity (logit models for contact with P and response | contact) -> ppbleib
10 Calibration of cross-section n+1 for individuals -> wqbap; wqmip; wqp

8.2.1 Design weights for the wave n households in the (n+1)th wave

New “household design weights” were generated for the (n+1)th wave from the cross-sectional weights for households of wave n, taking into account people moving into households from within Germany. This is done using a weight share procedure. Births, deaths or moves out of households have no influence on the weight; moves into households from within Germany, on the other hand, increase the inclusion probability of a household as the individuals who have moved into the household also had the chance of being included in the sample in all previous waves. Thus for the weighting, if individuals had moved into the household from within Germany, the previous inclusion probability was increased by the mean inclusion probability in the respective subsample (as it is not possible to reconstruct precisely what inclusion probability the new household members' households had in all previous waves). The new design weight for subsample i $dw_i hh_{n+1}$ is therefore calculated from the old cross-sectional weight $wq_i hh_{n+1}$:

$$\frac{1}{dw_i hh_{n+1}} = \frac{1}{wq_i hh_n + (n_{sample\ i}/n_{population\ i})}$$

The new design weight is only an intermediate step and is therefore not included in the data.

8.2.2 Design weights for the wave n+1 refreshment sample

From wave 2 on there is a yearly refreshment of the BA sample selecting only benefit units (Bedarfsgemeinschaften) in which no member was receiving benefits in July of the previous years. The refreshment sample is drawn in the same sampling points as the initial samples. From wave 5 on it is extended to the 97 additional sampling points of the replenishment samples. Analogous to the special pps procedure used to draw the first register data sample, which is described in Rudolph/Trappmann (2007), the sample size is proportional to the share of new benefit recipients in the population in the sampling point (at the time when the sampling points were selected). The calculation of the design weights is described in the same article. However, from wave 2 on, the number of benefit units in a household was no longer taken into account. The design weight of the refreshment sample is included in the variable *dw_ba* for all cases from refreshment samples.

8.2.3 Response propensity models for panel households

In this step, the probability of re-participating is estimated for each household which participated in the previous wave on the basis of logit models for willingness to participate in a panel, loss of contact and refusal. The models contain survey design features (e. g. mode, number of call attempts), aspects of the previous wave interview situation (e. g. item nonresponse or partial unit nonresponse), household/respondent characteristics (e. g. gender, age, education, country of birth, labour force status, house ownership, household size) and area characteristics (e. g. municipal size) as is state of the art in longitudinal studies (cf. Watson/Wooden 2009).

The predicted propensities of all three models are multiplied. The reciprocal value of this product can be found in the variable *hpbleib*. This variable serves a double purpose:

a. The longitudinal weight of a household for the period [wave_n; wave_n+k] between waves can then be calculated as the product of the cross-sectional weight for wave_n and the product of all *hpbleib* for wave n to wave n+k-1.

b. The product of the updated household design weight from step 1 (cf. section 8.2.1) multiplied by *hpbleib* (which we call “modified cross-sectional weight”) serves as a base for calculating a new cross-sectional weight for wave n+1.

Note that this procedure works only for households with monotonous drop-out patterns. Households who drop-out for one wave and return in the next wave cannot be treated this way. The treatment of those temporary dropouts is specified in section 8.2.5.

8.2.4 Non-response weighting for households from the wave n refreshment sample

For the households in the refreshment samples, non-response was modelled in a two-step procedure as was done for the first wave. The full lists of variables in the models and coefficients are described in the wave specific data reports cited above. The participation probability derived from this can be found in variable *prop_t0*.

8.2.5 Propensity models for temporary dropouts

From wave 3 on there are households in the PASS-dataset that have returned after temporariliy dropping out of the panel.[11] The longitudinal weights cannot be applied to this group of households which means that weighted longitudinal analyses can only be performed with the balanced panel of households who participated in all waves within the period considered for the longitudinal analysis. Allowing for non-monotonous patterns would result in an exponentially growing number of weights by wave (Lynn/Kaminska, 2010).

For temporary dropouts, first the probability of dropping out in wave n given participation in wave n-1 is derived from the propensity models for the transition from wave n-1 to wave n.[12] Then a simple propensity model (containing only final disposition code of the previous wave, mode, sample and whether it is a split-household) is specified predicting the probability of returning in wave n+1 given a dropout in wave n.

The reciprocal value of the product of the predicted probabilities of these two models is multiplied with the calibrated household weight of wave n-1 to calculate a modified cross-sectional weight which is used as a base for calculating a new cross-sectional weight for wave n+1.

8.2.6 Integration of weights by convex combination

The temporary dropouts are from the same population as that for which new base weights have been calculated in step 3 (8.2.3). Thus integrated weights can be calculated as a convex combination of the modified cross-sectional weights for the two subsamples (cf. Spiess/Rendtel, 2000). Formulae for this can be found in chapter 6 of Berg et al. (2011, 2013a,b,c).

8.2.7 Response propensity models for panel persons

The most important longitudinal weight is not the one at the household level but the one at the individual level, as the units here are stable over time. Participation propensities for individuals with monotonous dropout patterns are modelled in the same way as the model for households shown in step 3 (8.2.3). As the participation of the household is a precondition for the participation of the individual, the models contain similar variables. In addition, characteristics of the respective individual (e. g. age, item missings in the previous wave) are taken into account. The predicted propensities of the models are again multiplied.

[11] In PASS households which do not participate in two consecutive waves are no longer contacted.

[12] It can simply be calculated as 1-*hpbleib* for that wave.

The reciprocal value of this product can be found in variable *ppbleib*. The longitudinal weight of an individual for the period [wave_n;wave_n+k] between waves can then be calculated as the product of the cross-sectional weight for wave_n and all *ppbleib* for wave n to wave n+k-1. The full lists of variables in the models and coefficients are described in the field and method report of TNS Infratest for wave 2 and in the data report of infas for wave 3. Again temporary dropouts must be treated separately.

8.2.8 Integration of the weights to yield the total weight before calibration

This step involves combining the household weights of the latest refreshment sample and the panel households, which have been modified by the non-response modelling (steps 3 and 4) and the integration of temporary dropouts (step 6). The one-time integration of replenishment samples is described in section 8.3. The double selection probability of a newly sampled benefit recipient who was living in the same household as benefit recipients in the previous year but without being a member of the benefit unit him/herself is ignored. This is likely to be a rare population as four conditions have to be fulfilled simultaneously: (i) benefit recipiency at the reference date for sampling of a given wave, (ii) no benefit recipiency at all previous reference dates, (iii) currently living in a household with a former reference date benefit recipient, (iv) not having lived with such a person at any previous reference date. As the frames are disjunctive under this assumption, the weights of the register data sample alone remain unaffected by the integration of the refreshment sample (those of the general population sample on its own of course, too). The new design weights of the benefit recipient sample project in the cross-section to all individuals who were living in a household containing at least one benefit unit at one of the reference dates in either 7/2006, 7/2007, 7/2008, 7/2009, 7/2010 or 7/2011. It is only when calculating new weights for the total sample that it becomes necessary to adjust the weights for all households in receipt of benefits at prior reference dates. For this adjustment the inclusion probability in the respective other sample is estimated for cases from the general population sample and the refreshment samples. For cases from the refreshment sample, the mean selection probability in the general population sample in the respective postcode sector and the average participation probability in that sample are assumed. For cases from the general population sample, if they are (according to survey data) new recipients of Unemployment Benefit II who first received the benefit between the date of sampling for wave 1 and the date of sampling for one of the refreshments, the mean selection probability of a household in the refreshment sample in the respective postcode sector and the average participation probability in that sample are assumed. The two weights from 4 and 6 are then integrated to form a new total weight.

8.2.9 Calibration to the household weight, wave n+1, cross-section

The steps described above are followed by another calibration – of the weights from step 6. At the household level, raking (wave 3) or GREG (all other waves) is used to calibrate the weights to the benchmark statistics of the Federal Statistical Office for the respective year

(2007 in wave 2 to 2011 in wave 6) and for households in receipt of benefits the weights are adjusted to the statistics of the Federal Employment Agency for July of the respective year. The calibration process is described in detail in Kiesl (2010) for wave 1 and 2 and in the data reports of infas for wave 3 to 6 (Berg et al., 2011, 2013a,b,c).

8.2.10 Calibration to the person weight, wave n+1, cross-section

As in wave 1, the person weights were calibrated under the restriction that they differ as little as possible from the calibrated household weights. The calibration is therefore not based directly on the person weights of the previous wave. The calibration process is described in detail in Kiesl (2010) for wave 1 and 2 and in the data reports of infas for wave 3 to 6 (Berg et al., 2011, 2013a,b,c).

8.2.11 Estimating the BA cross-sectional weights for households and individuals not in receipt of Unemployment Benefit II

Finally, some households and individuals remain that can not be assigned a BA cross-sectional household weight or a BA cross-sectional person weight by means of calibration. They belong to one of the following three groups which did not receive benefits at any of the reference dates after wave 1 but which belong to the population of the BA sample (households with receipt of Unemployment Benefit II at the reference date in 7/2006, 7/2007, 7/2008, 7/2009, 7/2010 or 7/2011 and individuals in households with receipt of Unemployment Benefit II at the reference date in 7/2006, 7/2007, 7/2008, 7/2009, 7/2010 or 7/2011).

1. From the refreshment sample: individuals in the household who are not members of a benefit unit: here the person weight is obtained from the BA household weight of the respective wave after calibration (*wqbahh*) by dividing it by the proportion of these individuals who gave a personal or senior citizens' interview – provided that their household was participating.

2. Panel households in which nobody was in receipt of Unemployment Benefit II any longer at the reference date of the current wave: The household retains the BA weight before calibration (from step 8). Individuals in these households with interviews in both, the previous and the current wave are given a new BA person weight which is obtained by multiplying their BA person weight from the previous wave by the reciprocal re-participation probability *ppbleib*. Individuals in these households who did not provide a personal interview in the previous wave are given a new BA person weight calculated by dividing the BA household weight of their household for wave n+1 by the proportion of such individuals who participate provided that their household is taking part.

3. Individuals who are not members of a benefit unit in panel households that are still in receipt of Unemployment Benefit II at the current reference date: Individuals in these

households with interviews in both waves are given a new BA person weight which is obtained by multiplying their BA person weight of the previous wave by the reciprocal re-participation probability *ppbleib*.

8.3 Integration of the weights of the replenishment samples with the ongoing panel samples

The following section is adapted from for the purposes of this User Guide from the wave 5 data report by Berg et al. (2013b).

Integrating the replenishment samples with the ongoing panel samples in terms of weighting is not trivial since weights must be integrated several times. This integration, which only became necessary in wave 5, was performed between steps 7 (8.2.7) and 8 (8.2.8). In a first step cases in the ongoing panel that are not (or no longer) part of the population of the refreshments are identified. The weights of this subset are not affected by the replenishment samples. After this step, the remaining panel sample and the panel supplements are now two random samples of the same population with known inclusion probabilities. Thus, the concept of convex combination (Spiess/Rendtel, 2000) was applied to the fusion of these two samples and the combination with minimum variance was chosen.

The weights of the combined population samples are designed to project the initial general population sample (sampled from the Microm database) and the replenishment from municipal registers to all households in Germany. Thus, separate weights were calculated initially for the general population panel sample and the general population replenishment sample following the steps described in sections 8.1 and 8.2.1–8.2.7. Then, the general population panel was integrated with the general population replenishment (sample = 6) via a convex combination to obtain the population weight before calibration.

The weights of the combined BA samples are designed to project the original BA sample, the four refreshment samples with new entries to unemployment benefit II and the new BA replenishment sample to all households which received benefits at one of the reference days in July of the years 2006–2010. The subset still receiving unemployment benefit II at the reference date in July 2010, is to be projected to all recipients at that date. Initially, separate weights were calculated according to the steps described in sections 8.1 and 8.2.1–8.2.7 for the ongoing BA panel sample and the BA replenishment. The BA replenishment (sample = 7) then had to be integrated with recipients of UB II from all BA samples of waves 1–4 (convex combination). Cases from the BA samples from waves 1–4 which did no longer receive UB II maintain their weight in this case. This resulted in the new BA weight before calibration.

Since the new BA refreshment sample (sample = 8) and the BA panel samples are disjoint, all cases maintain their weights during the integration. A fusion of the samples for the calculation of the BA weight before calibration was not necessary.

Following this, the general population weights (wwmihh) and BA weights (wqbahh) were integrated to generate the total weights as described in section 8.2.8.

8.4 Datasets and variables

Like the individual and household datasets, the weighting datasets *hweights* (household weights) and *pweights* (person weights) are organised as long files.

The file *hweights* therefore now contains the following variables:

Table 26: Overview of the variables in the household weights data file (*hweights*)

Name	Label	Remarks
hnr	Household number (current)	Used together with *welle* for linking the datasets
welle	Indicator for survey wave	Used together with *hnr* for linking the datasets
sample	Subsample	Indicates whether BA or Microm weights are used
dw_mi	Design weight – Microm sample	Is the selection probability (during sampling) in the respective subsample (gross)
dw_ba	Design weight – BA sample	Is the selection probability (during sampling) in the respective subsample (gross)
dw	Design weight – total sample	Is the selection probability (during sampling) in the total sample (gross)
prop_t0	Participation probability in the sampling year of the subsample	Is the probability of the household taking part in the year when the subsample was drawn, as predicted by means of a logit model
wqhh	Projection factor – household (total)	Projection factor for the cross-section of the respective wave (total)
wqmihh	Projection factor – household (Microm)	Projection factor for the cross-section of the respective wave (Microm)
wqbahh	Projection factor – household (BA)	Projection factor for the cross-section of the respective wave (BA)
hpbleib	Reciprocal re-participation probability – household ($w_n \rightarrow w_{n+1}$)	Reciprocal value of the probability of the household participating in the survey again in the following wave, as predicted by means of a logit model

The file *pweights* contains the following variables:

Table 27: Overview of the variables in the person weights data file (*pweights*)

Name	Label	Remarks
pnr	Unchanging personal ID number	Used together with *welle* for linking the datasets
welle	Indicator for survey wave	Used together with *pnr* for linking the datasets
sample	Subsample	Indicates whether BA or Microm weights are used
wqp	Projection factor – person (total)	Projection factor for the cross-section of the respective wave (total)
wqmip	Projection factor – person (Microm)	Projection factor for the cross-section of the respective wave (Microm)
wqbap	Projection factor – person (BA)	Projection factor for the cross-section of the respective wave (BA)
ppbleib	Reciprocal re-participation probability – person ($w_n \rightarrow w_{n+1}$)	Reciprocal value of the probability of the individual participating in the survey again in the following wave, as predicted by means of a logit model

Part II

Using the PASS data

9 Key variables

Daniel Gebhardt and Arne Bethmann

Key variables are used to identify units and observations and to establish links between different datasets. These variables are essential whenever a certain research question requires information from different datasets which must therefore be combined before analyses can be carried out.

This section aims to explain the key variables of PASS and how they are put to use. In a first step, this section will explain how the key variables are connected to the structure of the scientific use file (*SUF*) and its datasets, which were discussed in section 5. Secondly, these variables are described in more detail. Also, an overview of the key variables included in the different datasets of the scientific use file is given. The section concludes with several practical examples illustrating the use of the key variables.

9.1 Key variables and their connection to the structure of the scientific use file

The structure of the *SUF* and its datasets were illustrated in chapter 5. There it was shown that the datasets of the *SUF* can be classified by their *level* (household or individual), their *type* (register, cross section, weight or spell) and which *formats* they are prepared (wide, long, spell) in. Which key variables can be used to identify units and their respective observations depends on the *level* and *format* of the dataset.

On the household as well as on the individual level PASS uses specific identification numbers (ID) that are constant across waves. These ID-numbers can be used to identify certain units – households or persons – in all datasets of the *SUF* and across all waves.

A certain household can be identified via the current household number *hnr* and can be related to its household of origin via the original household number *uhnr*.[13] Households keep their *hnr* across waves. If a part of an already surveyed household splits off, the newly formed household gets a new *hnr* and keeps it for future waves.

Individuals are assigned a constant personal ID-number *pnr* when they are a member of a successfully surveyed household in PASS for the first time. Persons keep their *pnr* across waves, even if they change between households, e. g. when they leave their household of origin and form a new split-off household.

Using only the ID-numbers – *hnr* on the household and *pnr* on the individual level – one can clearly identify a unit in each of the different datasets but not necessarily a certain observation. Additional information is required to clearly identify an observation, which depends on the format of the specific dataset in question.

[13] For households that have been drawn directly for one of the samples, the *uhnr* is identical to the *hnr*. Households that have split off from another household in PASS carry an *uhnr* representing the *hnr* of the household of origin.

Datasets that are prepared in *wide format* (the register datasets) contain only one observation per unit while the wave-specific information is stored in wave-specific variables, e. g. *age1* for a persons' age in wave 1, *age2* for the age in wave 2 and so on. In these datasets each unit has exactly one observation and therefore can be clearly identified using the ID-variables.

Datasets that are prepared in *long format* (the cross-sectional datasets and the weights) as well as the datasets that are prepared in *spell format* (the different spell datasets) can contain more than one observation per unit. Datasets in long format contain as many wave-specific observations for each unit as there are waves this unit was interviewed in, e. g. if a household was interviewed twice, the household dataset contains two observations for this household – one for each wave with an interview. Therefore, the wave indicator *welle* is required in addition to the household or personal ID-number in order to identify an observation clearly. In spell format datasets, the spell number *spellnr* has to be taken into account when identifying an observation. The spell datasets contain as many observations as there are episodes reported by the household or person, e. g. the employment spells contain two observations for a person if this person reported two episodes of employment.

All datasets include key variables which are used to identify units and observations and to establish links to other datasets of the *SUF*. The key variables included in the dataset are listed in tables 28 and 29. For further information about their meaning and on how to use them, see the corresponding chapter in Berg et al. (2013c). We strongly recommend PASS users to make themselves familiar with the structure of the datasets, their meaning and the key variables before combining different datasets.

9.2 Example: Merging household data with the individual dataset

If household data are to be merged with the individual dataset (e. g. the information on the type of the household which is contained in the variable *hhtyp*), then the two relevant key variables – the household number (*hnr*) and the wave indicator (*welle*) – must be used.

```
use PENDDAT.dta, clear
merge m:1 hnr welle using HHENDDAT.dta, keepusing(hhtyp)
tab _m welle
drop if _m == 2
```

The tabulation of the *_merge* variable shows that information from the household dataset was merged for some cases from wave 2 (N=140) and wave 3 (N=190) for which no personal interviews were available. These are re-interviewed households without personal interviews in the respective wave. These cases are dropped for the example.

9.3 Example: Merging the household weights with the household dataset

The household dataset and the household weights are available in the same format and on the same level. Accordingly, the datasets can be merged directly. The same procedure is used for merging the individual dataset and the person weights.

Table 28: Overview of key variables in the scientific use file

Key variable	Description
hnr	*Current household number* Eight-digit, constant ID number of a household, which is allocated when the household joins the panel. The first digit indicates the wave in which the household was first part of the gross sample of PASS. E. g.: 10010008 – household in gross sample for first time in 1st wave 21011685 – household in gross sample for first time in 2nd wave . . .
uhnr	*Original household number* Eight-digit, constant ID number that points to the original household. In the case of households that were drawn directly for one of the subsamples, the *uhnr* is the same as the respective *hnr*. In the case of households which have split off from panel households (split-off households) the *uhnr* corresponds to the *hnr* of the household from which the split-off household originated.
hnr$	*Household number in wave$* Eight-digit, constant ID number of the household in wave$ of PASS. This variable is only contained in the register datasets processed in wide format.
pnr	*Constant personal ID number* Ten-digit, constant ID number of the individual. The *pnr* is allocated when a person first joins a PASS survey household. The first eight figures consist of the household number of the household to which the person belonged when he or she joined PASS and the last two figures are the serial number that the person had within this household. E. g.: 1001000801 – person joined the PASS in household 10010008 and had the serial number 01 in this household
zplfd$	*Serial number of the target person in the household in wave$* Two-digit serial number within the household in wave$, which indicates the person's position in the household structure. Within a particular household the *zplfd* is constant in principle. If a person moves to a different household between the waves, then a new *zplfd* is allocated in the new household – in this case *zplfd1* and *zplfd2* differ. Serial numbers that were already used for a certain household in one of the previous waves are not allocated to anyone else. The numbering of new people in a household begins at N+1 (N = highest *zplfd* ever allocated in that household).
welle	*Indicator for survey wave* Both the household and individual datasets as well as the corresponding weighting files of PASS are processed in long format. For every interview that was conducted with a household or a person there is a row in the data matrix. By means of a wave indicator (*welle*) it is possible to assign these different observations for a household or a person to the respective survey wave.
spellnr	*Spell number* As in the datasets processed in long format, another variable is necessary in addition to the household and personal ID numbers in order to identify observations clearly in the spell datasets. In the different subject-related datasets the spells were put into chronological order and then each one was given a serial number, the spell number, within the household or the person. It is not easily possible to relate spell information clearly to a survey wave as the spells contain cross-wave information.

Table 29: Key variables in the datasets of the scientific use file

Dataset	Key variables contained						
	hnr	***uhnr***	***hnr$****[*]	***pnr***	***zplfd$***	***welle***	***spellnr***
Household level							
Household register (*hh_register*)	x	x	x				
Household dataset (*HHENDDAT*)	x	x				x	
Household weights (*hweights*)	x					x	
Household dataset on retirement provision (*HAVDAT*, wave 3 only)	x	x				x	
Unemployment Benefit II spells (*alg2_spells*)	x						x
Individual level							
Person register (*p_register*)		x	x	x	x		
Person dataset (*PENDDAT*)	x	x		x		x	
Children dataset (*KINDER*)	x	x		x		x	
Person weights (*pweights*)				x		x	
Factorial survey: job offer acceptance (*VIGDAT*, wave 5 only)				x			
Person dataset on retirement provision (*PAVDAT*, wave 3 only)				x		x	
Employment biographies (*bio_spells*, from wave 2)				x			x
One-Euro-Job spells (*ee_spells*, from wave 4)				x			x
Measure spells (*mn_spells*, wave 2 and 3 only)				x			x
Measure spells (*massnahmespells*, wave 1 only)				x			x
Unemployment Benefit I spells (*alg1_spells*, wave 1 only)				x			x

[*] "$" represents the number of a certain wave and indicates a wave-specific variable, e. g. *hnr*$ represents the household number in wave$ – therefore the variable name for wave 1 is *hnr1*.

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
tab _m welle
```

The tabulation of the *_merge* variable shows a perfect match of the household dataset and the household weights. For each household that was interviewed in a certain wave an observation from the weighting dataset was merged. See chapter 12 on the use of the weights.

9.4 Example: Merging information from the individual dataset with the person-specific spell data

When merging spell data and the household or individual dataset, it is always necessary to take the different logics of the datasets into account. Whilst the household and individual datasets contain wave-specific observations of the study units, the spells cannot be assigned clearly to one particular wave. A spell of employment, for example, can span several survey dates. This spell is then visible in the data structure as a single observation with its respective start and end dates. If, for instance, individual-level information is to be merged with the person-specific spell data (spells of employment, unemployment, gaps, employment and training measures), then these different data structures have to be taken into consideration. As it is not straight forward to assign every spell clearly to a particular survey wave, only the personal ID number can be used as a key variable. The information from the individual dataset therefore has to be converted to wide format first and then merged with all of a person's spells. This is demonstrated below using the example of the date of the personal interview which is available in the individual dataset and is to be merged with the employment spells.

First the individual dataset, reduced to the relevant variables, is converted to wide format. For this the information on the interview date, which has been stored in wave-specific observations so far, is restructured. Instead of there being one observation per survey wave, there is now only one single observation for each individual in the dataset. The information on the interview date is now stored in the wave-specific variables *pintdat1*, *pintdat2*, et cetera. For many individuals the spell dataset contains more than one observation. By linking via the personal ID number, the respective interview dates of each individual wave are added to each of a person's spells and are available for further calculations.

The biography spell dataset consists of different spell types: employment, unemployment, as well as other times out of employment, e. g. retirement, housewife/-husband, and military or civil service. You can keep certain types of spells by using the variable *spelltyp*. In the example only the employment spells are kept in the dataset.

```
use PENDDAT.dta, clear
keep pnr welle pintdat
reshape wide pintdat, i(pnr) j(welle)
```

```
la var pintdat1 "Datum des Personeninterviews in Welle 1"
la var pintdat2 "Datum des Personeninterviews in Welle 2"
la var pintdat3 "Datum des Personeninterviews in Welle 3"
la var pintdat4 "Datum des Personeninterviews in Welle 4"
la var pintdat5 "Datum des Personeninterviews in Welle 5"
la var pintdat6 "Datum des Personeninterviews in Welle 6"
save PINTDAT.dta
use bio_spells.dta
keep if spelltyp == 1
merge m:1 pnr using PINTDAT.dta
tab _m
drop if _m == 2
```

The tabulation of the *_merge* variable shows that no employment spell is available for over 19,000 individuals. Some of these individuals were only interviewed in the 1st wave, some had not reported any employment spells since and some were not asked about their employment owing to a filter. These cases are dropped.

10 Register data

Daniel Gebhardt and Arne Bethmann

In addition to the cross-sectional datasets at the household and the individual levels (*HHENDDAT* and *PENDDAT* respectively), the various spell datasets (*alg2_spells*, *bio_spells*, *ee_spells*) and the weighting datasets (*hweights*, *pweights*), the scientific use file of PASS also contains a household register dataset and a person register dataset (*hh_register*, *p_register*). In contrast to the other datasets, these two files are processed in wide format, i. e. there is exactly one observation available per household or individual. Information referring to individual survey waves is stored in wave-specific variables. The wave to which a piece of information refers is indicated by a counter at the end of the respective variable – thus the variable *alter1* in the person register, for example, contains the person's age in the 1st wave, and *alter2* is accordingly the person's age in the 2nd wave and so on. The register datasets are prepared in such a way that they can easily be converted from wide format to long format, for example using the "reshape" command in Stata. Subsequently the register information can be merged with the survey datasets, which are available in long format. Households which are not interviewed in certain waves, individuals in households which are not interviewed and individuals who no longer belong to a sample household in a later wave can be identified via the respective net variables.[14] In addition, in these cases the wave-specific household number (*hnr$*) is allocated the code "-6". In the following sections the structure and contents of the household register dataset and the person register dataset are presented and their use demonstrated using two examples.

[14] These are described later in this chapter.

10.1 Household register

All of the households which have been successfully surveyed at least once in the sense of PASS (see section 7.1 for the definition) are contained in the household register. Accordingly, households from the gross samples of the individual waves which were not successfully surveyed and households that have split off from panel households and have not been interviewed are not contained in the household register. In addition to the identifiers, the register dataset contains in particular wave-specific information on the survey status of the households (*hnettok$*, *hnettod$*), on the sample (*sample*), the sampling year (*jahrsamp*), the Unemployment Benefit II receipt of the household on the sampling date (*alg2samp*) and on the number of benefit units in the household. The household register therefore makes it possible to establish in which waves a household was interviewed in PASS and why no interview is available for certain waves. In this way a preliminary selection of households can be conducted – for example all of the households that were interviewed in all of the waves can be selected.

10.2 Person register

The person register dataset contains all individuals who were a member of a PASS survey household in at least one wave, irrespective of whether an interview at the individual level has already been conducted with them or not. In addition to the constant personal ID number as the identifier and details regarding e. g. the person's gender (*sex*) and wave-specific age (*alter$*), the person register dataset contains information about which household the person belonged to in the survey waves (*hnr$*) and what position he/she occupied in the structure of these households (*zplfd$*). The person register thus makes it possible to allocate individuals to households in specific waves. Furthermore, the person register dataset contains information regarding the individuals' survey status in the individual survey waves (*pnettok$*, *pnetto$1*), which can be used, for example, to identify fully surveyed households, to distinguish between reasons for non-response and to clarify people's whereabouts.

In addition to the person-related information, the person register dataset also contains information on the benefit unit to which the individual was assigned. These benefit units are so-called "synthetic" benefit units created on the basis of the current legal situation at the particular time and based on information about the ages of the households members and relationships between them, irrespective of whether they are currently receiving Unemployment Benefit II (see chapter 3.2). The information about the benefit units is available as wave-specific information. It must be taken into account that this information was generated each time on the basis of the information available for the individual waves. Via the benefit unit ID number (*bgnr$*) it is possible to identify the individuals who together constitute a benefit unit. Here it must be taken into consideration that new numbers are allocated in each wave and that there is no continuation in the longitudinal section. Furthermore, the dataset contains information on the type of benefit unit (*bgtyp$*) and on the benefit receipt of the benefit unit on the sampling date (*bgbezs$*) and the survey date of the current wave (*bgbezb$*).

The person register dataset also contains pointer variables referring to the mother living in the household (*zmhh$*), the father living in the household (*zvhh$*) and the partner living in the household (*zparthh$*). These pointers each contain the ten-digit personal ID number of the person who is the target person's mother/father/partner.

10.3 Example: Selection of the households that were successfully surveyed in the 1st and 2nd wave and were receiving Unemployment Benefit II on the sampling date

The net variables are available in two levels of detail – in a "short", single-digit variant (*hnettok1*, *hnettok2*, . . .) and a "detailed", two-digit variant (*hnettod1*, *hnettod2*, . . .). The two-digit net variables differentiate the single-digit codes further. The single-digit code "2" in *hnettok2* (household not successfully surveyed, only in gross sample) is further differentiated in *hnettod2* in the codes beginning with "2". This makes it possible to establish why the household could not be successfully surveyed in the 2nd wave, for example because the household could not be reached (*hnettod2=20*) or because it refused to participate (*hnettod2=21*). As only households that were successfully surveyed are to be selected here, the information in *hnettok1* and *hnettok2* is sufficient. After retaining only the cases that were successfully surveyed in both the 1st and the 2nd waves and were receiving Unemployment Benefit II on the sampling date, only the relevant variables (*hnr, alg2samp*) are retained, the dataset is sorted by household number, stored temporarily and merged with the observations from the first two waves of the household dataset, which has also been sorted according to *hnr*.

```
use hh_register.dta, clear
keep if hnettok1 == 1 & hnettok2 == 1 & alg2samp == 1
keep hnr alg2samp
save hh_register_vorb1.dta, replace
use HHENDDAT.dta, clear
keep if welle == 1 | welle == 2
merge m:1 hnr using hh_register_vorb1.dta
tab _merge alg2samp, m
```

An examination of the *_merge* variable indicates that 6210 observations (from 3105 households) from the household register dataset which were interviewed in both waves and were in receipt of Unemployment Benefit II on the sampling date were merged with the individual dataset.

10.4 Example: Identification of the personal interviews with the heads of households

The household register dataset contains the wave-specific information about which person the household interview was conducted with. In order to mark the personal interviews of these heads of households, it is first necessary to prepare the household register and convert it into long format. First of all, only the required variables are retained – the household number and the wave-specific pointer to the target persons of the household interview. Then the dataset is reshaped from wide format to long format. For this the household number serves as an ID variable that identifies an observation. In the course of the reshaping process a wave indicator (*welle*) is created which is needed for merging with the individual dataset.

However, before the register which has been converted into long format can be merged with the individual dataset, some observations have to be deleted. If a household was not interviewed in one wave, then the pointer variable referring to the head of the household was given the value "-6" (household not interviewed in wave or not in gross sample) for this wave. A household that was interviewed for the first time in the 2nd wave, for example in the context of the refreshment sample, has the value "-6" for the observation referring to the 1st wave. These observations cannot be merged with the individual dataset and should therefore be deleted. After this, the pointer variable *pnrzp* is renamed *pnr*, as the data is to be merged via the constant personal ID number. After the register dataset has been prepared, it is stored temporarily and merged with the individual dataset.

```
use hh_register.dta, clear
keep hnr pnrzp1 pnrzp2 pnrzp3
reshape long pnrzp@, i(hnr) j(welle)
drop if pnrzp == -6
ren pnrzp pnr
drop hnr
save hh_register_vorb2.dta, replace
use PENDDAT.dta, clear
merge 1:1 pnr welle using hh_register_vorb2.dta
tab _merge
drop if _merge == 2
gen hhvorst = _merge == 3
```

The tabulation of the *_merge* variable shows that in 1028 cases there is no personal interview available with the person who completed the household interview in that wave. As there is no information about them from personal interviews, these observations – which were merged from the person register – can be deleted. All of the cases for which the merging was successful (*_merge == 3*) were the head of the household in the particular wave and are flagged via the variable *hhvorst*.

11 Spell data

Daniel Gebhardt and Arne Bethmann

In all waves, the scientific use file of PASS included spell datasets on the household level as well as on the individual level (see table 30 for an overview). Whereas the dataset on

Unemployment Benefit II receipt of the household (*alg2_spells*) was continued in the 2nd wave, the survey concept for the other two spell datasets, Unemployment Benefit I receipt (*alg1_spells*) and participation in employment and training measures (*massnahmespells*), was thoroughly revised. In the course of this revision process, it was decided not to continue the data structure of the spell datasets on employment and training measures and Unemployment Benefit I receipt used in the 1st wave but to create new datasets (see chapter 4.4. and 4.5 in Gebhardt et al. 2009). Periods when the respondent received Unemployment Benefit I are surveyed from the 2nd wave onwards as part of the periods of registered unemployment in the biography module. For every period when the respondent was registered as unemployed, information is gathered as to whether he/she received Unemployment Benefit I and if so, from which start date and to which end date. Periods of Unemployment Benefit I receipt are therefore embedded in a period of registered unemployment and are no longer surveyed directly as they were in the 1st wave. The way in which participation in employment and training measures is surveyed was revised because it had emerged that in some cases it was not possible to identify the type of measure clearly with the concept used in the 1st wave. Starting with wave 2, the type of measure is identified right at the beginning of the module using a multiple choice question. Due to low numbers of reported spells for other training measures, only "One-Euro-Jobs" are recorded from wave 4 onwards (*ee_spells*).

Another important innovation regarding the spell datasets results from the fact that the concept for surveying periods of employment, unemployment and economic inactivity was altered in the 2nd wave. Instead of only asking about the status as of the interview date, as was done in the 1st wave, a biography module is used since wave 2 to record spells of employment and registered unemployment retrospectively for a certain period. In wave 2, episodes since January 2005 up to the date of the interview were recorded. In wave 3, persons who already answered questions about their employment and unemployment biography in wave 2 were asked about the period since the interview in wave 2. Persons who were not interviewed in wave 2 or were not asked about this topic, reported about the periods since January 2006 up to the date of the interview. In wave 2 as well as in wave 3, gaps as of the date of the interview date or periods of more than three months duration for which the respondent reported neither employment nor unemployment are caught by a gap module. If the gap existed due to the omission of a period of employment or unemployment or if the dating of a reported spell was incorrect, this could be corrected here. If the gap represented a time of economic inactivity this could also be specified in the gap module. Starting with wave 4, spells of employment, unemployment and economic inactivity are surveyed in chronological order in an integrated questionnaire module. These spell data are provided as a single dataset in the scientific use file (*bio_spells*).

The period of time covered by a spell dataset differs between households and between persons. The beginning of the period depends on the wave in which the respective module in the questionnaire was first asked of the household/person and additional characteristics. With each wave, the year starting from which the respondents were asked to report episodes was increased by one year to keep the length of the first retrospective period constant.[15]

[15] E. g. in wave 2 the employment spells were recorded for the first time in the personal interviews. The

Table 30: Overview of the spell datasets in the scientific use

Dataset	Contents	Data collection in waves
Household level		
alg2_spells	Information on periods when the household received Unemployment Benefit II; periods of cuts in Unemployment Benefit II	since wave 1 (dataset covers period from January 2005 – date of interview in the most recent panel wave)
Individual level		
bio_spells	Information on periods when the respondent was ▪ employed with a monthly income of more than 400 € ▪ registered as unemployed or was participating in a employment or training measure run by the Employment Agency ▪ not in employment (and not registered as unemployed)	since wave 2 (dataset covers period from January 2005 – date of interview in the most recent panel wave)
ee_spells	Information on periods when the respondent was participating in a “One-Euro-Job”	since wave 4 (dataset covers period from January 2009 – date of interview in the most recent panel wave)
Datasets which are not continued		
mn_spells	Information on periods when the respondent was participating in an employment or training measure	waves 2 and 3 only (dataset covers period from January 2006 – date of interview in wave 3)
massnahmespells	Information on periods when the respondent was participating in an employment or training measure	wave 1 only (dataset covers period from January 2005 – date of interview in wave 1)
alg1_spells	Information on periods when the respondent was receiving Unemployment Benefit I	wave 1 only (dataset covers period from January 2005 – date of interview in wave 1)

The beginning of the covered period depends not only on the wave of the first interview but also on additional characteristics, e. g. if there was a later change in the household composition[16] or when the person who answered the household questionnaire in the last interview moved out.[17]

The end of the period depends on the wave in which the respective module was last asked of the household/person. If a household/person missed a wave (temporary drop-out), the resulting gap in the spells was filled in the next interview if the household/person had been asked the respective module before. If a person was not asked a certain module due to a filter, the resulting gap was not necessarily filled in the next interview. Before using the spell datasets it is reasonable to take a look into the questionnaires and to trace the way the spells were recorded. This will help to interpret times were no spell data is available for a household/person.

The spell datasets of PASS have a comparable structure. In addition to an identifier (household or personal ID number), they also contain a spell number, which numbers the individual spells within a household (*alg2_spells*) or a person (*bio_spells, ee_spells, mn_spells, massnahmespells, alg1_spells*) consecutively in chronological order and makes it possible to identify them clearly together with the household or personal ID number. Furthermore, generated date variables for the beginning (*bmonat*, *bjahr*) and the end (*emonat*, *ejahr*) of the respective spell can be found in the datasets. These variables were recoded (e. g. information on seasons was recoded into definite months) and cleansed (e. g. missing codes were set for implausible values). In addition, if these variables contained censored spells, the interview date was entered for the end of the spell. In contrast, the date variables as they were reported by the respondent (e. g. *BIO0200, BIO0300, BIO0400, BIO0500* in the *bio_spells*), which are also included, were not altered.[18] Following content-related information on the various spell types, all of the spell datasets contain a censoring indicator (*zensiert*) for spells that were still ongoing on the respective last interview date, in other words, right-censored spells. Generated variables (e. g. ISCO-88 coding of occupational activities) can be found at the end of each list of variables in the spell datasets.

Finally, some important peculiarities of the spell data in PASS should be pointed out. Due to the orientation towards actual spells here, it is generally not easy to relate the spells to specific waves, as spells may span more than one survey date. Furthermore, observations are not available for all households or individuals in the spell data. This may be the case if there were no relevant spells or if the corresponding questions were not asked due to the filters.

For identifying individual spells, the identifier variable (*hnr* or *pnr*) and the spell number are

respondents were asked to report episodes since January 2005. In wave 3 this date was altered to January 2006.

[16] In this case, UBII-episodes in the first interview were asked since the date of the last change of the household composition.

[17] In this case the former household was asked for episodes of UBII-recipiency since the move-out while the new household (split-off household) of this person was asked for episodes since the interview date of the former household.

[18] Exceptions to this are the merging of two spells and the spells of Unemployment Benefit II receipt surveyed in the first wave.

always required for a clear selection, as there are often several observations available per household or person. This also has to be taken into account when linking spell data and the household and individual datasets. As several spells are frequently available and there is also no wave indicator for the individual observations in the spell data, a wave-specific reference is not possible without further work.

A spell can include several pieces of information of the same kind that refer to different points in time. These are recorded in individual variables within the same observation in the spell dataset (e. g. the amount of benefits the household received: *AL20800* if the information was recorded in wave 1, *AL20801* for wave 2 etc.). As long as a reported episode has not ended, the information from the last interview always corresponds to that interview date. However, if an episode has ended, the information from the last interview corresponds to the reported end date. If there are several pieces of information recorded in different waves, the ones which were reported while the episode had not ended correspond to the respective interview date. If there is no information recorded for an episode in a given wave the respective variable is assigned the missing code "-9". The wave a given variable in the spell refers to, can be read from the variable labels.

The following example demonstrates the generation of a variable containing the latest information about the amount of benefits received per month for each Unemployment Benefit II spell. Variables for the other cross-sectional information can be generated in the same way.

11.1 Example: Using the cross-sectional information included in the spell datasets

First a new variable is created, *hoehebez*, which is assigned code -3 (not applicable), as details about the amount of benefit received are only available for Unemployment Benefit II spells that were still ongoing at the interview date in at least one wave. Then the generated variable is filled with the information from *AL20800* (amount of benefit received per month in wave 1), *AL20801* (amount of benefit received per month in wave 2) and the respective variables for the following waves using a loop. Information is only incorporated into *hoehebez*, however, when it does not involve the values "-3" (not applicable) or "-9" (item not surveyed in wave). A cross-sectional variable on the amount of benefit received is given the value "-3" if information about the spell was gathered in the respective wave (new details surveyed or previous details updated) but the spell was not ongoing on the interview date. The variable is assigned the code "-9" if no information was collected about this spell in the respective wave. First *hoehebez* is filled with the information on the amount of benefits received as recorde in the cross-sectional variable for wave 1 (*AL20800*) and then, in the consecutive loop runs, is replaced by the values of the cross-sectional variables referring to the respective waves. Thus *hoehebez* is replaced with the latest available information for this spell.

```
use alg2_spells.dta, clear
gen hoehebez = -3
forvalues i = 0/5 {
  replace hoehebez = AL2080‘i’ if AL2080‘i’ ~= -3 & AL2080‘i’ ~= -9
}
```

12 Weights

Mark Trappmann

12.1 Recommendations for the use of *surveyset* in Stata

All of the weights in PASS are so-called *probability weights*: the weight of a household or a person is equivalent to the reciprocal value of its/his/her inclusion probability (adjusted by non-response modelling and calibration). In Stata, starting with version 9, *probability weights* have to be set using the *surveyset* command (see Stata Corp 2007). However, *surveyset* not only has the purpose of defining the weights to be used, but also of defining the aspects of the survey design that have an impact on the standard errors.

There are two different options for doing this in Stata: by specifying the design or by using replication weights. In the first option, the aspects of the survey design that influence the standard error have to be entered in the command line. Besides the weights, these aspects are clusters, stratification characteristics and finite population corrections for sampling without replacement. The effect of calibration on the standard error and other factors such as pps-sampling cannot be taken into account. The second option, on the other hand, makes use of a set of replication weights, which are calculated for all units of the study using processes such as jackknifing, BRR or bootstrapping. These procedures also potentially permit the calibration to be taken into account.

There are no replication weights available for PASS to date, so researchers will have to use the first variant for the *surveyset* for PASS. However, the complex sample design of PASS cannot be used for variance estimation with the *surveyset* command in all details. We recommend the following approach:

```
svyset psu [pw=wqX], strata(strpsu)
```

Here *wq*X stands for the adequate weight for the intended analyses. An indicator for the *primary sampling units* (which are the same for both subsamples) is the variable *psu* in the household dataset *HHENDDAT*. The strata for the selection of the primary sampling units are represented by the variable *strpsu* in the same dataset. Strata with fewer than two units in the sample were collapsed. In the case of sampling with replacement, neglecting strata and clusters from the second level onwards (in PASS these would be households) only leads to an “... unimportant underestimation ...” (Särndal/Swensson/Wretman, 1992: 144 pp.). If the sampling rate is very low, the variance estimation for sampling without replacement can be approximated very well using the formulae for sampling with replacement. This is the case for PASS (only approximately 3.6% of the postcodes in Germany were selected for the survey). There is therefore no need to indicate finite population corrections or further clusters (here: households). However, the recommended *surveyset* then takes neither calibration nor pps-sampling into account, nor the low finite population correction for sampling without replacement. The resulting standard errors are too large and thus should be considered conservative estimates.

From wave 2 onwards there are rare cases where strata defined by the variable *strpsu* only contain one single primary sampling unit because all of the respondents in the other PSU belonging to the stratum have dropped out. When a stratum consists of only one PSU, Stata cannot estimate standard errors. The easiest way to circumvent this problem is to retain the cases from all waves – even if only one wave is being analysed – and to declare the wave of interest to be a subpopulation using the *subpop* option of the survey commands (see Stata Corp 2007: 53 pp.).

If one works solely with the dataset of a single wave after wave instead (e. g. with a refreshment sample), Stata, from Version 10 onwards, provides various approximation procedures for cases of strata with only one PSU (the *singleunit* option of the *svyset* command, see Stata Corp 2007), but none of them solve the problem entirely satisfactorily.

Singleunit certainly assumes that the single PSU in the sample is also the only one in the population and that the variance between PSUs in this stratum is therefore zero. As there are several PSUs in every stratum in the population of PASS, the basic assumption is not correct. This setting thus results in the variance being underestimated.

In the case of *singleunit scaled*, the stratum with missing variance is assumed to have a variance equal to the mean variance in the other strata. As these are rather small strata, however, the variance is likely to be larger in reality.

With *singleunit centered* a variance within the stratum with only one PSU is estimated by assuming that the (unknown) stratum mean is equal to the grand mean. The variance of the stratum is then estimated from the mean of the single PSU in the stratum and the grand mean.

Another general remark is in place on this subject: restrictions to subpopulations using "if" or "keep if" can hinder the estimation of standard errors if the restriction results in strata with only one PSU. We recommend to implement restrictions using "*subpop*" and not "if" or "keep if". The only exception is the restriction to one of the PASS subsamples. Here the restriction with "if" is appropriate. Examples are given in the next section.

12.2 Use of the cross-sectional weights

In this section examples are given on how to use the cross-sectional weights for different purposes. For all examples code in Stata 12.1 is given. All Stata code is printed in separate lines in `Courier New` and can be copied from this User Guide and pasted right into your Stata do-file-editor. Please, replace

```
"[PATH_TO_DIRECTORY_OF_ORIGINAL_PASS_DATA]"
```

by the name of the path where the original PASS data are on your computer and replace

```
"[PATH_TO_DIRECTORY_FOR_WEIGHTING_EXERCISES]"
```

by the name of the path where you want to store the results of this training session. In case you are using any later version of Stata than version 12.1 all you have to do in order to ensure getting the same results is precede the code by

```
version 12.1
```

All of the cross-sectional weights are projection factors. Dividing these weights by their mean value results in weights that add up to the sample size. Design weights (*dw_mi*, *dw_ba*, *dw*) and the estimated propensities propensities for the initial wave (*prop_t0*) are provided in PASS; however, we recommend using the calibrated weights. Researchers who nevertheless wish to do without calibration should bear in mind that although division of the household weights by the adequate participation propensities estimated for the respective subsample does yield modified household design weights, these weights cannot simply be transferred to all responding persons in the households as they do not take into account person non-response within participating households (partial unit-nonrepsonse). Use of design weights at the person level thus aditionally requires an estimation of the person's participation propensity, given participation of the household.

The following sections provide examples showing how to use the cross-sectional weights for different research questions.

12.2.1 Analyses of benefit recipients in July 2006

The most efficient way to obtain findings on the population of the BA sample in the 1st wave (households in which there was at least one benefit unit receiving benefits in accordance with SGB II as of July 2006; referred to below as "households receiving benefits in July 2006") is to use only the BA sample and the relevant weights. Proceeding in this way is more efficient than using the total sample, as the weights in the BA sample have less variance. Furthermore, the analyses have to be restricted to *sample==1*, as cases from the refreshment samples are otherwise taken into account too.

Analyses at the household level To make analyses of households receiving benefits in July 2006, researchers should use *wqbahh*. The example below demonstrates its use in Stata 12.1. It is intended to calculate the number or percentage of households receiving benefits which are in possession of a car (variable *HLS0800a*). To start with, the household weights have to be merged with the household dataset, then the *surveyset* command has to be carried out, and then the projected value can be calculated:

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
svyset psu [pw=wqbahh], strata(strpsu)
svy, subpop(if welle==1): tab HLS0800a if sample==1, ///
count cell format(%9.0g)
svy, subpop(if welle==1): tab HLS0800a if sample==1, ///
cell ci format(%9.0g)
```

Approximately 37.9% of the households receiving benefits in July 2006 had a car at the time of the survey in the 1st wave, 62.1% did not have a car, and the percentage with no valid response is extremely low. Whilst the first tabulation command shows the projected number and percentages of individuals with and without a car, the second tabulation gives the percentage and the corresponding 95% confidence intervals with the option “ci”. The confidence interval is [36,0; 39,7]. It would also be possible to dispense with the restriction if *sample==1* as the weight *wqbahh* in wave 1 is only defined for the cases from sample 1 (BA register data sample as of the reference date in July 2006).

The values for the number and percentage of car owners in the same population at the time of the survey in the 6th wave in the relevant population are obtained as follows:

```
svy, subpop(if welle==6): tab HLS0800a if sample==1, ///
count cell format(%9.0g)
```

Approximately 52.4% of the households receiving benefits in July 2006 had a car at the time of the interview in the 6th wave.[19] The fact that the value increased compared with that of the first wave is likely to be associated with the fact that a considerable number of these households have managed to end benefit recipiency between the first and sixth wave. If researchers are solely interested in those households that are still in receipt of benefits at the time of the most recent interview, then the command has to be restricted to this set. As it is not a separate sample, a restriction with “if” would result in an underestimation of the variances in this case. The restriction is to be carried out using *subpop* (see Stata Corp 2007: 53 pp.). The information as to whether a household is receiving benefits on the survey date is contained in the variable *alg2abez* in *HHENDDAT*. Here the value 1 means that the household was drawing benefits, the value 2 means that it was not in receipt of benefits and -5 means that it is not possible to establish from the information available whether the household was receiving benefits. The command is therefore:

```
svy, subpop(if alg2abez==1 & welle==6): tab HLS0800a if ///
sample==1, count cell format(%9.0g)
```

Of the households which were receiving benefits in July 2006 and were also still in receipt of benefits at the survey date in the 6th wave, only 28.0% have a car. This value has

[19] Households that had split off from wave-1 households since then by moving out are included in this analysis.

pronouncedly decreased compared to the first wave. It is highly plausible that deprivation has increased for this selective group that is still or again on benefits 5 to 6 years after the reference date for sampling. The corresponding confidence intervals are displayed using the option “ci”.

```
svy, subpop(if alg2abez==1 & welle==6): tab HLS0800a if ///
sample==1, cell ci format(%9.0g)
```

[24.3%; 31.8%] is reported as the 95% confidence interval. This confidence interval lies entirely outside the corresponding interval for 2006. If the 6th wave was selected using an if condition instead of the *subpop* option, in other words by entering the following command:

```
svy, subpop(if alg2abez==1): tab HLS0800a if welle==6 & ///
sample==1, cell ci format(%9.0g)
```

then the message “Note: missing standard errors because of stratum with single sampling unit” would appear. As described above there are three approximation procedures for this case available which can be applied modifying the *svyset* command:

```
svyset psu [pw=wqbahh], strata(strpsu) singleunit(certainty)
svy, subpop(if alg2abez==1): tab HLS0800a if welle==6 & ///
sample==1, cell ci format(%9.0g)
```

produces anti-conservative estimates, i. e. smaller confidence intervals. However, in this case results only differ only on the second position after the decimal point (when the proportion is displayed as a percentage).

```
svyset psu [pw=wqbahh], strata(strpsu) singleunit(scaled)
svy, subpop(if alg2abez==1): tab HLS0800a if welle==6 & ///
sample==1, cell ci format(%9.0g)
```

in this case produces slightly larger confidence intervals [24.2%; 31.9%] than the more exact estimation using *subpop*. Finally,

```
svyset psu [pw=wqbahh], strata(strpsu) singleunit(centered)
svy, subpop(if alg2abez==1): tab HLS0800a if welle==6 & ///
sample==1, cell ci format(%9.0g)
```

again leads to results that only differ from the on the second position after the decimal point from the more exact procedure. These examples indicate that when using PASS - due to a small number of strata with only one PSU - differences between the estimation procedures are negligible.

Analyses at the benefit unit level Researchers working on recipiency of Unemployment Benefit II are often not interested in households but in benefit units. If the above question on the percentage of households receiving benefits in July 2006 which are in possession of a car is to be transferred to benefit units, the PASS data can be used to answer the question as to how many benefit units live in a household that has a car (as the benefit units were identified retrospectively, there are no questions in the questionnaire relating directly to benefit units - it is therefore not possible to identify which benefit unit owns the car in a household consisting of several benefit units). This question is relatively easy to answer, using the variable *nbgbezug*, which states how many benefit units in joint receipt of Unemployment Benefit II a household contains as of the sampling date. [20] The fastest way to do this is to multiply the household weights by this value.

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
gen bgweight=wqbahh*nbgbezug
svyset psu [pw=bgweight], strata(strpsu)
svy, subpop(if welle==1): tab HLS0800a if sample==1, ///
count cell format(%9.0g)
```

The percentages deviate slightly from those in the analysis presented above (37.9% of households receiving benefits, but 38.2% of the benefit units receiving benefits had a car in their household in wave 1). Above all, however, the absolute numbers are different: the sum of all households receiving benefits was 3,882,013, whereas the sum of all benefit units receiving benefits is 4,011,889, and matches the BA benchmark statistics due to the calibration.

In contrast, with PASS it is not possible to calculate the percentage of car owners as of the survey date of the 2nd or any later wave for the benefit units of the first wave. As the compositions of benefit units are constantly changing due to deaths, births, moves into and out of the household, and also due to members reaching certain age limits (25 and 65 years of age), this kind of analysis across waves should be conducted at the level of more stable units.

Analyses at the person level Analyses at the person level are similarly simple. The person dataset *PENDDAT.dta* and the weight wqbap should be used in this case. An intermediate step becomes necessary, as the variables *psu* and *strpsu* are only contained in the household dataset. The following example estimates the number of individuals aged 15 and above[21] in households receiving benefits who have a migration background (variable *migration*).

[20] For this variable, the decisions required when the statements do not clearly identify how many benefit units are receiving Unemployment Benefit II in the household were made in the same way as for the calibration process. Every user is of course free to make his or her own decisions on the basis of the Unemployment Benefit II spells.

[21] As younger people are not interviewed in person, the PASS data can only be used to establish characteristics about them which are surveyed in the household questionnaires (e. g. age, gender). The household weights should be used in this case.

```
use HHENDDAT.dta, clear
keep hnr welle psu strpsu
save psuinfo, replace
use PENDDAT.dta, clear
merge 1:1 pnr welle using pweights.dta
drop _m
merge m:1 hnr welle using psuinfo
svyset psu [pw=wqbap], strata(strpsu)
svy, subpop(if welle==1): tab migration, ///
count cell format(%9.0g)
```

According to this calculation, about 59.7% do not have a migration background, 24.0% migrated to Germany themselves, at least one parent migrated to Germany for a further 7.6%, and at least one grandparent for another 1.8%. The code "Item not surveyed in questionnaire"[22] applies to 3.6%. This is due to the fact that the data from the short questionnaire for people aged 65 and above are stored in the same dataset as data from the standard personal questionnaire. People aged 65 and above are assigned this code for questions that are not asked in the senior citizens' questionnaire. In order to run analyses excluding these individuals, researchers can limit the frequency count to data from the standard questionnaires (*fb_vers=1*)

```
svy, subpop(if welle==1 & fb_vers==1): tab migration, ///
count cell format(%9.0g)
```

In much the same way as shown above for households, the analyses for individuals from households receiving benefits in July 2006 can be run for the survey date of later waves (*e. g. welle==6*) and restricted to those people who were still living in a household in receipt of benefits on the survey date in a later wave (*e. g. welle==6* & *alg2abez==1*).

The person weights of the BA sample project to all individuals in households receiving benefits. Some households, however, consist of several synthetic benefit units, not all of which receive benefits. Researchers wishing to project only to persons who are members of benefit units according to Social Code Book II have to exclude individuals who did not belong to a benefit unit on the sampling date. The variable *bgbezs1* from the dataset *p_register* provides information on a person's affiliation with a benefit unit in receipt of benefits at the sampling date for wave 1:

```
drop _m
merge m:1 pnr using p_register.dta
keep if pnetto1==2 | pnetto1==3
svy, subpop(if bgbezs1==1 & fb_vers==1 & welle==1): ///
tab migration, count cell format(%9.0g)
```

[22] For a further 1.2%, the variable cannot be formed due to missing information.

The percentage of individuals who migrated to Germany themselves is therefore marginally higher among the people who are members of a benefit unit, at 25. 1%, than among people living in a household receiving benefits (24.8%).

12.2.2 Analyses on the resident population of Germany

Analyses on the resident population of Germany can be carried out both with the total weights and with the population sample weights. In most cases the results will differ only slightly. The percentage of households with a car in the total population (in wave 1 in this case) is therefore estimated either with the following commands using the total weights:

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
svyset psu [pw=wqhh], strata(strpsu)
svy, subpop(if welle==1): tab HLS0800a, cell ci format(%10.0g)
```

or alternatively with the population sample weights:

```
svyset psu [pw=wqmihh], strata(strpsu)
svy, subpop(if welle==1): tab HLS0800a, cell ci format(%10.0g)
```

In the first case, the percentage of households with a car is 75.9% (95% confidence interval of 73.9% to 77.7%), and in the second case 75.6% (95% confidence interval of 73.5% to 77.6%). The confidence interval is slightly narrower when the total weights are used, as in this case the part of the population receiving benefits under SGB II is represented much more precisely, which is why we prefer to use these weights. The same applies to the person weights.

12.2.3 Analyses on benefit recipients at different points in time

Section 12.2.1 explained how the data can be projected onto the total population of the BA register data sample of the 1st wave (households with at least one benefit unit that was in receipt of benefits in accordance with Social Code Book II in July 2006). As a result of its design, however, PASS is more flexible and makes it possible in principle to make projections onto the benefit recipients at any point in time since the benefit was introduced in January 2005.

Analyses on benefit recipients in July 2011 PASS takes a first step in this direction with the annual refreshment samples of the register data sample. The refreshment samples (samples 3, 4, 5, 8, 9) consist of households in which there was at least one benefit unit receiving benefits at the reference date in July of the respective wave but of which no member was living in a household with at least one benefit unit in receipt of benefits at any previous reference date. When all BA samples are taken together, the result is a sample made up of households receiving benefits at any of the reference dates in July 2006, July 2007, July 2008, July 2009, July 2010 or July 2011 - admittedly an unusual population. However, if this combined population is restricted to households that were also still in receipt of benefits in accordance with Social Code Book II at the reference date for the most recent wave (currently 2011 for wave 6), then these cases can be projected to all households with Unemployment Benefit II recipiency at this last reference date. The annual refreshment of the sample thus enables us to remain "representative" for the benefit recipients in July of the previous year using the integrated benefit recipient samples.

The indicator for benefit recipiency as of the sampling date (of the respective wave) at the household level is the variable *alg2abez* in *HHENDDAT*, which is available for each household in every wave. At the person level it is the variable *bgbezs** in *p_register*. Here * is a placeholder for the respective wave (*bgbezs1* in wave 1, *bgbezs2* in wave 2 and so on).

We take up the examples from section 12.2.1 again in the following when we calculate the percentage of households with a car and the percentage of individuals with a migration background as of the interview date of the 6nd wave; but restricted this time to all benefit recipients as of July 2011. We do not have to tell Stata which samples to use as the BA sample weights are defined only for cases from the BA samples.

Households in receipt of benefits in July 2011

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
svyset psu [pw=wqbahh], strata(strpsu)
svy, subpop(if alg2abez==1 & welle==6): tab HLS0800a, cell ///
ci format(%9.0g)
```

30.5% of all households in receipt of benefits in July 2011 had a car on the interview date of the 6th wave. A 95% confidence interval of 28.2% to 32.9% is obtained.

Individuals in receipt of benefits in July 2011

```
use PENDDAT.dta, clear
merge 1:1 pnr welle using pweights.dta
drop _m
merge m:1 hnr welle using psuinfo
drop _m
merge m:1 pnr using p_register.dta
svyset psu [pw=wqbap], strata(strpsu)
svy, subpop(if bgbezs6==1 & welle==6 & fb_vers==1): ///
tab migration, count cell format(%9.0g)
```

Of all the individuals in receipt of benefits in accordance with Social Code Book II in July 2011, 25.6% migrated to Germany themselves, a further 8.7% have at least one parent who migrated to Germany and another 2.0% have at least one grandparent who migrated.

Analyses on benefit recipients at the latest interview date When working with the original BA sample only (*sample==1*) and the appropriate weights, the results refer to recipients in July 2006. For analyses of this population, this approach achieves the greatest statistical power, as the BA weights have a relatively low variance. However, researchers will wish to carry out many analyses – especially on fast-changing characteristics – using the latest available data, to which many characteristics refer, such as employment status, income or working hours. The survey date of the first wave is between 6 and 13 months after the sampling date; that of the second wave is even 18 to 25 months later. When working on the latest available data exclusively with the BA sample, researchers can only make statements about so-called "stayers", those who continued to receive benefits from the sampling date until the survey date. In view of a considerable turnover (37% of people receiving benefits under SGB II in January 2005 were no longer doing so by December 2006 (Graf, 2007)), this group may differ significantly in its makeup from the current benefit recipients. The refreshment of the benefit recipient sample cannot solve this problem. It can be solved, however, by merging the benefit recipient sample with the population sample. The price for this is, however, a substantial loss of statistical power.

Analyses of benefit recipients at the latest interview date on the household level
Representative results for current benefit recipients can therefore only be obtained using the total weights. The variable for whether the household is currently receiving benefits (*alg2abez*) is contained in the household dataset (*HHENDDAT*). Estimations are therefore relatively simple for analyses at the household level. The example below shows this, again using the question of car ownership.

```
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
svyset psu [pw=wqhh], strata(strpsu)
svy, subpop(if alg2abez==1 & welle==6): tab HLS0800a, ///
cell ci format(%9.0g)
```

Of the households currently receiving benefits, 31.9% had a car on the survey date of the 6th wave. If this were estimated using the BA weights and the BA sample,

```
svyset psu [pw=wqba], strata(strpsu)
svy, subpop(if alg2abez==1 & welle==6): tab HLS0800a, ///
cell ci format(%9.0g)
```

a proportion of 30.5% would be estimated. However, as these data only include 'stayers', in other words households that were receiving benefits both on one of the sampling dates in July 2006-2011 and on the survey date, it is plausible that fewer of these households have cars than those that started receiving benefits between the last reference date and the survey date.

One consequence of using the total weights rather than the BA weights is the substantial increase in the confidence intervals. The variance of the total weights is significantly larger due to the very different sampling rates in the two subsamples. The analyses on car ownership in households receiving Unemployment Benefit II in July 2011, for which we can only work with the BA register data sample, result in a 95% confidence interval of 28.2% to 32.9%. For the survey date, we obtain a substantially larger 95% confidence interval of 28.6% to 35.5%.

Analyses on benefit recipients at the latest interview date at the benefit unit level
In comparison to the analyses referring to the sampling date in the previous section, an additional step has to be taken as there is no variable equivalent to *nbgbezug* for recipiency of benefits as of the survey date. This variable first has to be generated using the variable *bgbezb6* in *p_register*, which indicates for each benefit unit whether this particular community was receiving Unemployment Benefit II on the survey date of wave 6.[23]

```
use p_register.dta, clear
collapse (mean) hnr6 bgbezb6, by(bgnr6)
recode bgbezb6 (-5=0)
by hnr6, sort: egen nbgbezak=sum(bgbezb6)
collapse nbgbezak, by(hnr6)
rename hnr6 hnr
save hnr_nbgbezak.dta, replace
use HHENDDAT.dta, clear
merge 1:1 hnr welle using hweights.dta
drop _m
keep if welle==6
merge hnr using hnr_nbgbezak.dta
gen bgw_akt=wqhh*nbgbezak
svyset psu [pw=bgw_akt], strata(strpsu)
svy, subpop(if alg2abez==1): tab HLS0800a, cell ci format(%9.0g)
```

The estimated value of 31.8% differs only marginally from that obtained in the analysis at the household level. However, the value no longer refers to a sub-population of just over 3,185,000 households as in the section above, but to just over 3,212,000 benefit units receiving benefits as of the survey date. During the survey period, the number of benefit units varied between 3,288,000 (September 2012) and 3,377,000 (March 2012), according to the BA statistics. This benchmark value is thus not quite reached. The underreporting arises from the fact that, unlike in the figures referring to the sampling date, information on benefit recipiency at the time of the survey is not available from the register data for all

[23] In the sample code, “`recode bgbezb6 (-5=0)`” is used to treat all benefit units for which current recipiency of benefits is unclear on the basis of the survey data as non-recipients.

respondents. Thus the underreporting of benefit recipiency[24] using the latest available data is not corrected by means of calibration.

Analyses on benefit recipients at the latest interview date at the person level
Analyses can be transferred to the person level in much the same way as was done when using data referring to the sampling date. To start with, the person weights and the information for the *surveyset* should again be merged with the individual dataset. For analyses on individuals from households currently receiving benefits, the frequency counts should be limited to individuals with *alg2abez==1*. This variable has to be merged from the household dataset.

```
use HHENDDAT.dta, clear
keep hnr welle psu strpsu alg2abez
save psu_alg2_info, replace
use PENDDAT.dta, clear
merge 1:1 pnr welle using pweights.dta
drop _m
merge 1:m hnr welle using psu_alg2_info
drop _m
svyset psu [pw=wqp], strata(strpsu)
svy, subpop(if alg2abez==1 & welle==6 & fb_vers==1): ///
tab migration, count cell format(%9.0g)
```

According to this, of the individuals in households currently receiving Unemployment Benefit II, 57.8% have no migration background, 28.1% migrated to Germany themselves, 8.6% have at least one parent who migrated and 2.8% one grandparent who migrated. In most cases, however, analyses will not be limited to individuals in households receiving benefits, but to individuals in benefit units receiving benefits. This characteristic is contained in the person register. The following series of commands produces the percentage of migrants among individuals in benefit units aged between 15 and 64.

```
drop if welle<6
merge 1:1 pnr using p_register.dta
svy, subpop(if bgbezb6==1 & fb_vers==1): tab migration, ///
count cell format(%9.0g)
```

Analyses on benefit recipients at other points in time The biographical data on Unemployment Benefit II recipiency at the household level also make it possible in principle to perform analyses referring to other points in time which are between the sampling date and the date when the first wave of the survey was administered. However, variables such as *bgbezs1*, *bgbezb1* or *nbgbezug* are only provided for the two dates described above.

[24] As recipiency of Unemployment Benefit II is a socially undesirable characteristic, a certain amount of underreporting is not surprising. Compare Kreuter/Müller/Trappmann (2010) for a discussion of this underreporting.

Users who would like to run projections referring to other points in time will therefore have to generate analogous variables themselves. When doing this, both imprecision and the problem of benefit recipiency being under-reported will always have to be taken into account.

12.2.4 Comparison of benefit recipients with the general population

The large variety of options for studying benefit recipients and their households or benefit units benefit units shown above results in an equally large variety of options for comparing benefit recipients with the general population. Table 31 provides an overview. The total weights are to be used in all cases.

12.3 Use of the longitudinal weights

The basic principle of the longitudinal weighting is simple: Estimated reciprocal re-participation probabilities *hpbleib* and *ppbleib* are used for longitudinal weighting of households and persons respectively. The longitudinal weight for a household (or an individual) for the longitudinal section from wave 1 to wave 2 is obtained by multiplying the cross-sectional weight of the household (or person) for wave 1 by the reciprocal re-participation probability. The reciprocal re-participation probability is provided in the dataset hweights.dta (or pweights.dta) for all households (and persons) that participated in a given wave and in the subsequent wave. A large variety of different applications of longitudinal weights can be imagined depending on restrictions to certain waves, subsamples or cases with certain characteristics or analyses at different levels (household, benefit unit, individual). We would like to demonstrate the use of the longitudinal weights for some typical applications.

12.3.1 Persons of the resident population

One possible research question involving the longitudinal section could be how many persons (from the age of 15) of the resident population reported greater satisfaction with their standard of living in wave 2 than they did in wave 1 (variable *PA0300).* The population for such a question is all individuals who belonged to the resident population of Germany in wave 1 and wave 2.

Some preparations have to be made before, but they can also be used for the subsequent analyses. In a first step, wave 1 and the variables *psu* and *strpsu* are extracted from the household dataset

```
use HHENDDAT.dta, clear
keep hnr welle psu strpsu
keep if welle==1
drop welle
save psu_strpsu_w1.dta, replace
```

Table 31: Variables and their possible usage for comparing SGB II benefit recipients with the general population

Variable	Dataset	Values	Suitable for comparing ...
sample	*PENDDAT* *HHENDDAT*	1 BA sample 2 Microm sample 3 BA refreshment sample wave 2 4 BA refreshment sample wave 3 5 BA refreshment sample wave 4 6 general population replenishment sample 7 BA replenishment sample 8 BA refreshment sample wave 5 9 BA refreshment sample wave 6	households/persons in households with Unemployment Benefit II receipt at a certain reference date to households/persons of the general population and to households/persons in households who are new entries to benefit receipt.
alg2samp	*hh_register*	0 no benefit recipiency 1 benefit recipiency 2 no benefit recipiency acc. to survey (BA-sample) 3 benefit recipiency unclear acc. to survey (BA-sample) 4 benefit recipiency unclear acc. to survey (Microm-sample)	households/persons in households with Unemployment Benefit II receipt at a certain reference date for sampling to households/persons with no benefit receipt at the same reference date. Users may choose how to deal with cases that were receiving Unemployment Benefit II according to register information but not according to the survey.
bgbezs1 *bgbezs2* *bgbezs3* *bgbezs4* *bgbezs5* *bgbezs6*	*p_register*	1 UB II recipiency as of sampling date 0 no UB II recipiency as of sampling date	Persons in benefit units receiving Unemployment Benefit II at a reference date for sampling to persons in hypothetical benefit units not receiving Unemployment Benefit II at the same reference date. As this variable was used for the weighting process, a decision was made for every unclear case.
alg2abez	*HHENDDAT*	1 HH currently receiving UB II 2 HH currently not receiving UB II -5 No generation poss. (missing values)	a) Households/persons in households receiving Unemployment Benefit II on the survey date of a wave (*alg2abez=1*) to households/persons in households not receiving Unemployment Benefit II at the survey date of the respective wave.
bgbezb1 *bgbezb2* *bgbezb3* *bgbezb4* *bgbezb5* *bgbezb6*	*p_register*	1 UB II recipiency in wave 1 0 No UB-II recipiency in wave 1 -5 No generation poss. (missing values)	Persons in benefit units receiving Unemployment Benefit II on the survey date of a wave to persons in benefit units not receiving Unemployment Benefit II on the survey date of that wave

In a second step the weights from the first wave and the re-participation probabilities from wave 1 to wave 2 are stored.

```
use pweights.dta, clear
keep if welle==1
save pweights1.dta, replace
```

Now the individual dataset is retrieved. We have decided to run the analyses in wide format and therefore have to re-sort the dataset so that the variables *PA03001* (satisfaction with the standard of living in wave 1) and *PA03002* (satisfaction with the standard of living in wave 2) are retrieved. We only retain the variables that we require later.

```
use PENDDAT.dta, clear
keep pnr hnr welle PA0300
reshape wide PA0300 hnr, i(pnr) j(welle)
```

Now the three datasets are merged

```
rename hnr1 hnr
merge m:1 hnr using psu_strpsu_w1.dta
keep if _m==3
drop _m
merge m:1 pnr using pweights1.dta
drop _m
```

In order to make the tables clearer, a variable is created that indicates the relative level of satisfaction in wave 2 compared with wave 1.

```
gen rel_zufr=2 if PA03002>PA03001 & PA03001>=0 & PA03002>=0
replace rel_zufr=1 if PA03002==PA03001 & PA03001>=0 & PA03002>=0
replace rel_zufr=0 if PA03002<PA03001 & PA03001>=0 & PA03002>=0
replace rel_zufr=-1 if PA03001<0 | PA03002<0
label define rel_zufr_lb ///
2 "W2 zufriedener als W1" ///
1 "W1 und W2 gleich zufrieden" ///
0 "W2 weniger zufrieden als W1" ///
-1 "in mind. 1 Welle keine Angabe"
label values rel_zufr rel_zufr_lb
```

Finally the longitudinal weight is constructed and the weighted analysis follows

```
gen wp1_2=wqp*ppbleib
svyset psu [pw=wp1_2], strata(strpsu)
svy: tab rel_zufr, count cell format(%10.0g)
```

It refers to 67.7 million individuals who were at least 15 years old in wave 1 and were still resident in Germany on the survey date in wave 2. Of this group 34.6% were less satisfied in wave 2 than they were in wave 1. In contrast, 32.1% were more satisfied. For 33.1% the assessment had not changed.

12.3.2 Individuals in households receiving Unemployment Benefit II in July 2006

Now the same question can also be asked for the individuals in the benefit recipient sample of the first wave. How satisfied are these individuals in wave 2 compared with wave 1? The only difference to the previous analysis is that the BA weight has to be used instead of the total weight.

```
gen wbap1_2=wqbap*ppbleib
svyset psu [pw=wbap1_2], strata(strpsu)
svy: tab rel_zufr, count cell format(%10.0g)
```

Here 33.7% are less satisfied than in the previous wave, whereas 42.7% are more satisfied. The result refers to 6 ,045,000 individuals from the age of 15 who were living in a household which was receiving benefits in July 2006 and belonged to the resident population (irrespective of receipt) in wave 2. In this respect it is not surprising the majority is more satisfied than in wave 1, as some of them should have managed to leave benefit recipiency in the meantime. Researchers will therefore perhaps be more interested in how the satisfaction levels changed for those people who were receiving benefits on both survey dates.

12.3.3 Individuals in receipt of Unemployment Benefit II on two subsequent waves

As was the case in the analyses described above, for the question as to changes in the satisfaction levels of people who are still in receipt of benefits, the variables that indicate benefit recipiency on the survey date are required. These variables are contained in the person register, which is merged here.

```
merge m:1 pnr using p_register.dta
keep if _m==3
svyset psu [pw=wp1_2], strata(strpsu)
svy, subpop(if bgbezb2==1 & bgbezb1==1): tab rel_zufr, ///
count cell format(%10.0g)
```

As for the total population, a relatively balanced picture emerges here: 38.8% with increased satisfaction face 36.4% with a reduction in satisfaction. This preliminary work now also makes it possible to analyse rapidly the change in the satisfaction levels of people entering and leaving benefit recipiency.

```
svy, subpop(if bgbezb2==0 & bgbezb1==1): tab rel_zufr, ///
count cell format(%10.0g)
svy, subpop(if bgbezb2==1 & bgbezb1==0): tab rel_zufr, ///
count cell format(%10.0g)
```

Of the individuals leaving benefit recipiency, 55.0% are more satisfied, but 27.7% are less satisfied; of the individuals entering benefit recipiency, 46.4% are less satisfied, but 26.8% are more satisfied. This of course leads to the question as to whether the relatively large proportions of people who are less satisfied than they were in the previous year despite leaving benefit recipiency or are more satisfied despite entering benefit recipiency are associated with the fact that their income has hardly changed. This would go too far here, however.

12.3.4 Longitudinal weighting at the household level

First we present a simple example and then we address some of its problematic aspects. We answer the question as to how many households of the resident population acquired or gave up a car between wave 1 and wave 2. We use the same procedure as in the example for individuals described above: first, the dataset is created.

```
use hweights.dta, clear
keep if welle==1
save hweights1.dta, replace
use HHENDDAT.dta, clear
keep hnr uhnr welle HLS0800a psu strpsu
reshape wide HLS0800a psu strpsu, i(hnr) j(welle)
/* gen split=1 if hnr!=uhnr
replace hnr=uhnr if uhnr!=hnr
by hnr, sort: egen psu1x= mean(psu1)
replace psu1=psu1x if psu1==.
by hnr, sort: egen strpsu1x= mean(strpsu1)
replace strpsu1=strpsu1x if strpsu1==.
by hnr, sort: egen HLS0800a1x= mean(HLS0800a1)
replace HLS0800a1=HLS0800a1x if HLS0800a1==. */
merge m:1 hnr using hweights1.dta
keep if _m==3
drop _m
```

Then a variable is generated which expresses the change with regard to car ownership.

```
gen auto_neu=3 if HLS0800a1==2 & HLS0800a2==1
replace auto_neu=2 if HLS0800a1==1 & HLS0800a2==1
replace auto_neu=1 if HLS0800a1==2 & HLS0800a2==2
replace auto_neu=0 if HLS0800a1==1 & HLS0800a2==2
replace auto_neu=-1 if HLS0800a1<0 | HLS0800a2<0
label define auto_neu_lb ///
3 "Auto angeschafft" ///
2 "Auto behalten" ///
1 "weiterhin kein Auto" ///
0 "Auto abgeschafft" ///
-1 "in mind. 1 Welle keine Angabe"
label values auto_neu auto_neu_lb
```

Finally the weight is constructed and the table produced.

```
gen wh1_2=wqhh*hpbleib
svyset psu1 [pw=wh1_2], strata(strpsu1)
svy: tab auto_neu, count cell format(%10.0g)
```

1.7% of the households gave up a car, 2.2% acquired one, 76.2% kept one, 19.7% still do not have one. Instead of again distinguishing now between households receiving benefits and those not receiving benefits, we wish to discuss something more fundamental here:

The result produced above applies to all households of the resident population at the end of 2006 and their successor households existing as of the survey date in wave 2. As households are not units that are stable over time, a longitudinal analysis of households always requires a definition of what is to be regarded as the successor of a household in cases where the household composition changes. If the estimation is done as in this example, then the rules applied by PASS when allocating household numbers are used:

a) If individuals move into a household, the household number does not change. The new, larger household is the successor of the smaller household from the previous wave.

b) If household members die or move abroad, the household number does not change. The new, smaller household is the successor of the larger household from the previous wave.

c) If parts of the old household form a new household within Germany, then the household that retains the household number (and is therefore defined as the successor household) is the one that is reached via the original contact information (depending on the field this is either the telephone number or the address) or – if this does not apply to either of the new households – the one that is reached first. This household is defined as the successor here. This means that households which have split off from original households are not included in the analyses above. This could be one explanation for the finding that there are more households which have acquired a car than households that no longer have one: Households that were merged were included in the analysis. In contrast, of the households that have split up, only the half which remained at the old address or which was reached first was counted.

It is possible to incorporate split-off households into the longitudinal analysis, too. For this each split-off household has to be allocated the cross-sectional weight of the original household from wave 1 and a re-participation probability. The eight lines which are kept from being executed above (starting with /* and ending with */) have to be included for this. For that, please execute the code above again deleting both /* and */. The split-off households are projected to about 200,000 additional households and increase the percentage of households that had a car in wave 1 but no longer had one in wave 2 to 1.9%.

12.3.5 Longitudinal analyses utilizing more than two waves

All the examples above used only two subsequent PASS waves. In many applications more than two waves are used. For a longitudinal survey of n waves, 2^n-1 different combinations of waves are possible (cf. Lynn and Kaminska 2010). It is impossible for survey data producers to supply weights for all of these combinations. PASS longitudinal weights thus refer to the balanced panel only, i. e. to all those households who participated in all waves between the first and last wave of an analysis. Thus, if a researcher uses data from waves 3 to 6, a longitudinal weight can only be constructed for those respondents who continually participated in wave 3, 4, 5 and 6. If a researcher only uses waves 3 and 6 (e. g. because she analyses the extended health module) the same applies. Those who did not participate in wave 4 or wave 5 cannot be included in a weighted longitudinal analysis of waves 3 and 6. This implies a loss of power. Thus a methodologically advanced researcher could try to generate a more general longitudinal weight for her specific longitudinal analysis by estimating a participation probability of a person in wave 6 conditioned on participation in wave 3 and multiplying its reciprocal value with the wave 3 cross-sectional weight. The specification of the PASS propensity models is documented in the wave specific data reports (Berg et al., 2011, 2013a,b,c) and did not change much from wave 3. Most variables are taken from the previous (observed) wave and are available to users from the Scientific Use Files. Additional variables from the paradata of the survey (like number of contact attempts) can be supplied to users upon request.

13 Generated variables

Daniel Gebhardt and Arne Bethmann

The datasets of the scientific use file (*SUF*) of PASS include different types of variables. This section focuses on the generated variables which were created during the data editing process. They are meant to provide users a quick start or information that could not be included directly in the datasets of the scientific use file, e. g. information on the relationships between the household members. Detailed information about the generated variables can be found in the wave-specific data reports, e. g. an overview of the variables generated for a certain wave or the source variables they are based on (e. g. see chapter 4 in Berg et al. 2013c for wave 6). This chapter of the user guide will give a general introduction to the different types of generated variables and some notes on their use.

The datasets of the *SUF* contain six different types of generated variables:

- Variables generated due to coding of open-ended survey questions
- Variables generated due to dependent interviewing
- Constant characteristics
- Simple generated variables
- Theory-based construct variables

13.1 Coding of responses to open-ended survey questions

Some items of the survey were gathered as closed items with an open residual category or as open-ended items. In such cases, additional variables were usually generated which differed from the original variable only insofar as the information from the open-ended responses was coded to the corresponding categories where possible. Moreover, in some cases, new categories were created on the basis of the information from open-ended questions. The naming of these additional variables differs from that of the original variable in the last digit only, where the "0" was replaced by a "1". The items on country of birth, nationality, and the parents'/grandparents' country of residence before migration were also anonymised and given eloquent variable names.[25]

Information about the variables generated during the coding of open-ended survey questions in the different waves can be found in the wave-specific data reports (see e. g. chapter 4.1 in Berg et al. 2013c for wave 6).

13.2 Variables generated due to dependent interviewing

In various parts of both the household and the person interviews, information was gathered depending on responses given in previous waves. Information from the last interview was used in filter conditions, to display alternative texts or displayed directly in the current interview. Two primary objectives were pursued with the use of information from previous waves. First, in some modules only the changes since the last interview should be recorded depending on whether information on a certain topic was already recorded in a previous wave. In these cases, information from previous waves was used in filter conditions.

Second, in some parts of the interview the respondent was provided with information from previous waves. Therefore, the date of the last interview was displayed as part of the question text to narrow down the reference period. In other cases, particularly where episodes were updated, answers given in the last interview were integrated directly in the wording of a question to remind the respondent of the statements in the last interview. In doing so, the reporting of changes that did not really happen in the reference period should be prevented. These kinds of changes would be artifacts that result from recall errors or imprecise reports.

[25] *ogebland* (country of birth), *ostaatan* (nationality), *ozulanda–f* (parents'/grandparents' country of residence before migration)

Due to the use of dependent interviewing, the information for certain households/persons in the datasets can be incomplete if only a certain wave-specific observation is considered as it may only reflect the changes since the last interview. On the other hand, the information of a certain observation can also be complete up to the time of the interview if the household/person was interviewed for the first time about the topic in question.

In the course of data editing, the changes between two waves were combined with information from previous waves to provide generated variables with complete information for the cross-sectional datasets (*HHENDDAT*, *PENDDAT*) although only changes since the last interview were reported in the interview due to dependent interviewing. New or updated episodes since the last interview were used to update the respective spell datasets.

Detailed information on how information that was recorded using dependent interviewing was combined with information from previous waves can be found in the wave-specific data reports (see e. g. chapters 4.3, 5.6, 5.7 and 5.8 in Berg et al. 2013c for wave 6).

The so called *constant characteristics* (see section 14) are to be distinguished from this type of generated variable as it is assumed that they do not change over time. Therefore, they are only surveyed once for each household/person although corrections in a later wave are possible.

13.3 Simple generated variables

This type of variable covers, for example, variables for which different items of one construct that were surveyed separately for technical reasons were aggregated or for which information from the current wave was combined with information from the previous wave (such as the highest educational qualification) or for which important information was merged from other partial datasets (e. g. indicators for current receipt of Unemployment Benefit I or Unemployment Benefit II).

For households/persons that were asked for the first time regarding a certain topic, the simple generated variables can be created using only the information from this wave.

For households/persons that were already asked in the past regarding a certain topic, the simple generated variables can be distinguished by the origin of the source information for their creation. The three different types of simple generated variables are displayed in table 32.

Detailed information on the variables generated in the different waves and their respective source variables can be found in the wave-specific data reports (see e. g. chapter 4.4 in Berg et al. 2013c for wave 6).

13.4 Theory-based construct variables

Theory-based construct variables are variables whose generation requires more extensive re-coding and/or coding. In most cases, these variables have been empirically tested

Table 32: Types of simple generated variables in the cross sectional datasets (*HHENDDAT*, *PENDDAT*) for household/persons that were already asked in the past regarding a certain topic

Type	**Source variables for generation from**		**Description**
	wave of household's/ person's first interview regarding the topic	current wave	
constant (uv)	yes	no	In general, information from the first interview regarding the topic was carried forward – except for cases where falsely entered data was corrected in the current wave. E.g: *zpsex* (Gender of target person)
updated (fs)	yes	yes	The latest information from the previous wave was updated with the information recorded in the current wave E.g: *schul1* (Highest general school qualification)
independent (*neu*)	no	yes	In each wave, the variable was newly generated based only on the information that was recorded in this wave. E.g: *hhincome* (Household income per month)

Table 33: Information on constant characteristics – gender

Variable	Description	Dataset	Filled in for wave(s) of the first and repeated interview(s)
HD0100a–o	Gender of individuals 1 to 15 in the household	*HHENDDAT*	Yes, if person lived in household
zpsex	Gender of target person	*PENDDAT*	Yes
sex	Gender of target person	*p_register* *KINDER*	Information not wave-specific but contains the respective last correction

elsewhere and have a foundation in theoretical concepts. Moreover, some of them are standardized instruments used in social sciences or economics. Examples of such standardized instruments are the European Socio-economic Classification (ESeC), the International Standard Classification of Education (ISCED) or the equivalized household income. Detailed information on these variables in the different waves can be found in the wave-specific data reports (see e. g. chapter 4.5 in Berg et al. 2013c for wave 6).

14 Constant characteristics

Daniel Gebhardt and Arne Bethmann

Variables which are assumed not to change over time are only surveyed once in PASS. However, despite the constant nature of the characteristics in reality, changes in these variables are sometimes possible since, for example, incorrect entries may be corrected in subsequent interviews (e. g. in the case of gender). The following sections provide a brief overview of the constant characteristics that are available in PASS. The intention here is to show the conditions under which the variable was surveyed for the first time and to indicate the dataset in which it can be found. The key variables are disregarded here.

14.1 Gender

Information on a person's sex is gathered at the household level, either when the household in which the individual is living is first interviewed in the context of PASS or when the individual joins a sample household as a new member (e. g. when new individuals move into the household). In re-interviewed households, the interviewers had the opportunity to correct details regarding gender which had been recorded incorrectly in the previous wave. During the plausibility checks of the household structure, too, changes were occasionally made to the gender variables in households that attracted attention as a result of implausible relationships between the household members. Here, the gender data was checked on the basis of the first names. No retrospective changes of the data collected in earlier waves were made in either the household or the individual dataset.

14.2 Half year of birth

A person's half-year of birth was generated from the date of birth reported in the personal interview. Although it is a constant characteristic, the date of birth is asked for in every

Table 34: Information on constant characteristics – half-year of birth

Variable	Description	Dataset	Filled in for wave of the first interview	Filled in for wave(s) of repeated interviews
gebhalbj	Target person's half-year of birth, generated	*PENDDAT*	Yes	Yes

personal interview conducted. Among other things it serves to check whether the correct person is being interviewed. For re-interviewed persons the interviewers had the opportunity to correct details which had been entered incorrectly in the previous wave. If the half-year of birth differs from that in the previous wave as a result of the date of birth being corrected in the personal interview, this was understood as the correction of an incorrect entry. No retrospective changes were made to the information collected in the previous wave.

14.3 Migration background

A person's migration background is also understood as a constant characteristic and is only surveyed in the personal questionnaire in the first interview conducted with a person. The information on nationality (*PMI0400*, *PMI0500*), on temporary residence permits (*PMI0600*) and the type of residence/settlement permit (*PMI0650*), on the other hand, is gathered in every wave as changeable characteristics. In the senior citizens' interviews of the 1st wave no information was collected about whether the respondent's parents and/or grandparents migrated to Germany, and if so from where they migrated. It was therefore not possible to establish the migration background for senior citizens in the same way as in the standard personal interviews because information was only available about whether the respondent him/herself was born outside Germany. From the 2nd wave onwards the migration of parents and grandparents and their respective countries of origin are also surveyed in the senior citizens' interviews. In the first repeated interview after wave 1, all senior citizens were asked the questions. In subsequent waves this information will also be surveyed in the senior citizens' questionnaire in the first interview only.

The respondent's country of birth and information about the countries from which his/her individual parents and grandparents migrated to Germany are also made available in generated variables in which the information that was collected once is also taken over into subsequent waves. These variables are shown in Table 36. Furthermore, for a wave in which a person was not interviewed for the first time, the generated variable *migration* (see description e. g. for wave 6 in chapter 4 of Berg et al. 2013c) contains the information as to whether this person has a migration background and if so what this background is. Starting with wave 4 respondents are also asked for their first language (*PMI1200*) on the first interview.

14.4 Parents' education, vocational training, occupational status and activity

In wave 1, individuals whose mother and/or father did not live in the same household were asked about their parents' highest school qualification and their vocational qualifications. If

Table 35: Information on constant characteristics – migration background

Variable	Description	Dataset	Filled in for wave of the first interview	Filled in for wave(s) of repeated interviews
PMI0100	Target person born in Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI0200	Target person's country of birth, if not Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI0300a–b	Date of migration to Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI0700	Parents/grandparents born outside Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI0800a–f	Which parent/grandparent not born in Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI0900a–f	Which parent/grandparent migrated to Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI1000a–f	Country from which parent/ grandparent migrated to Germany?	*PENDDAT*	Yes	No (Except the first repeated interview for senior citizens first interviewed in wave 1)
PMI1200	Is German your mother tongue?	*PENDDAT*	Yes	No

Table 36: Information on constant characteristics – generated variables on migration background

Variable	Description	Dataset	Filled in for wave of the first interview	Filled in for wave(s) of repeated interviews
ogebland	Target person's country of birth, if not Germany? (incl. responses to open-ended questions, categorised)	*PENDDAT*	Yes	Yes
ozulanda–f[a]	Country from which parent/ grandparent migrated to Germany? (incl. responses to open-ended questions, categorised)	*PENDDAT*	Yes (Not surveyed for senior citizens in wave 1)	Yes
migration[b]	Target person's migration background (generated)	*PENDDAT*	Yes (Not generated for senior citizens in wave 1)	Yes

[a] The country from which the parents/grandparents migrated to Germany was surveyed for senior citizens for the first time in the 2nd wave.

[b] Not generated for senior citizens' interviews.

Table 37: Information on constant characteristics – social origin

Variable	Description	Dataset	Filled in for wave of the first interview	Filled in for wave(s) of repeated interviews
PSH0200	Target person's mother's highest general school qualification	*PENDDAT*	Yes	No
PSH0300a–i	Target person's mother's vocational qualifications	*PENDDAT*	Yes	No
PSH0310–PSH0380	Mother's occup. status and type of occup. activity when target person was aged 15	*PENDDAT*	Yes (Not surveyed in wave 1)	No (Except the first repeated interview for persons first interviewed in wave 1)
PSH0500	Target person's father's highest general school qualification	*PENDDAT*	Yes	No
PSH0600a–i	Target person's father's vocational qualifications	*PENDDAT*	Yes	No
PSH0610–PSH0680	Father's occup. status and type of occup. activity when target person was aged 15	*PENDDAT*	Yes (Not surveyed in wave 1)	No (Except the first repeated interview for persons first interviewed in wave 1)

the mother or father was living in the household, the information they provided in their own personal interviews was assigned to the target person. For individuals interviewed for the first time after wave 1, the parents' highest school qualifications and vocational qualifications were recorded as proxy information irrespective of whether the mother and/or father was living in the same household. Details about the qualifications which the parents may have given in their own personal interviews were thus no longer assigned to the children living in the household. People who had already been interviewed in the previous wave were not asked questions on this topic again. Furthermore, in wave 2 additional questions were incorporated about the mother and father's occupational status and occupational activity at the time when the target person him/herself was 15 years old. This information is also collected only once.

After the first interview, however, information is also available on the mother's school and vocational qualifications. It is contained in the generated variables *mschul1* / *mschul2* (mother's highest general school qualification without / with coding of responses to open-ended survey questions) and *mberuf1* / *mberuf2* (mother's highest vocational qualification without / with coding of responses to open-ended survey questions). Corresponding information for the target person's father can be found in *vschul1* / *vschul2* and *vberuf1* / *vberuf2*. The information on mother's and father's occupational status, which was first gathered in wave 2, are available in *mstib* and *vstib* in the individual dataset, also as generated variables. The generated variables cited here are described in the list of variables in the wave-specific data reports. Moreover, the information on the parents' occupational activity at the time when the target person was 15 years old was coded by Gesis-ZUMA (*misco*; *visco*) according to the 1988 International Standard Classification of Occupations (ISCO-88) published by the International Labour Office (ILO).

Table 38: Information on constant characteristics – sample information

Variable	Description	Dataset	Filled in for wave of the first interview	Filled in for wave(s) of repeated interviews
sample	Sample indicator	*HHENDDAT* *PENDDAT* *hh_register* *p_register* *hweights* *pweights*	Information not wave-specific	
jahrsamp	Sampling year	*hh_register*	Information not wave-specific	
alg2samp	Receipt of Unemployment Benefit II of the household on sampling date	*hh_register*	Information not wave-specific	

14.5 Sample indicator, sampling year and receipt of Unemployment Benefit II of the household on the sampling date

The sample indicator (*sample*), the sampling year (*jahrsamp*) and the receipt of Unemployment Benefit II of a household on the sampling date (*alg2samp*) are constant characteristics of the household which are defined once when the household joins the PASS sample. Individuals are assigned the sample indicator (*sample*) of the household to which they belong when they first become part of the PASS sample. Households which have split off from households already surveyed in the previous wave and are now surveyed as separate households in PASS take over the values of their original household in the variables *sample*, *jahrsamp* and *alg2samp*.

References

Achatz, Juliane; Hirseland, Andreas; Promberger, Markus (2007): IAB-Panelbefragung von Haushalten im Niedrigeinkommensbereich. Entwurf für ein Rahmenkonzept. In: Promberger, Markus (Ed.) Neue Daten für die Sozialstaatsforschung: Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung", No. 12/2007 in IAB-Forschungsbericht, Nürnberg, p. 11–32.

Bachteler, Tobias (2008): Dokumentation Record Linkage IEB-PASS. Tech. Rep., Institut für Soziologie, Universität Duisburg-Essen, unpublished.

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Bethmann, Arne; Fuchs, Benjamin; Gebhardt, Daniel (2013a): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Datenreport Wave 4. FDZ-Datenreport 08/2011 (EN), Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Bethmann, Arne; Fuchs, Benjamin; Trappmann, Mark; Wurdack, Anja (2013b): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Datenreport Wave 5. FDZ-Datenreport 06/2012 (EN), Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Bethmann, Arne; Fuchs, Benjamin; Trappmann, Mark; Wurdack, Anja (2013c): Codebuch und Dokumentation des "Panel Arbeitsmarkt und soziale Sicherung" (PASS). Datenreport Welle 6. FDZ-Datenreport 06/2013, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Marwinski, Karen; Gebhardt, Daniel; Wenzig, Claudia; Wetzel, Martin (2011): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Datenreport Wave 3. FDZ-Datenreport 06/2010 (EN), Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Beste, Jonas (2011): Selektivitätsprozesse bei der Verknüpfung von Befragungs- mit Prozessdaten. Record Linkage mit Daten des Panels "Arbeitsmarkt und soziale Sicherung" und administrativen Daten der Bundesagentur für Arbeit. FDZ-Methodenreport 09/2011, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Beste, Jonas; Eggs, Johannes; Gebhardt, Daniel; Gundert, Stefanie; Hess, Doris; Jesske, Birgit; Quandt, Sylvia; Trappmann, Mark; Wenzig, Claudia (2011): IAB-Haushaltspanel Lebensqualität und soziale Sicherung. Interviewerhandbuch Welle 5. FDZ-Methodenreport 03/2011, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Büngeler, Kathrin; Gensicke, Miriam; Hartmann, Josef; Jäckle, Robert; Tschersich, Nikolai (2010): IAB-Haushaltspanel im Niedrigeinkommensbereich Welle 3 (2008/09). Methoden- und Feldbericht. FDZ-Methodenreport 10/2010, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Büngeler, Kathrin; Gensicke, Miriam; Hartmann, Josef; Jäckle, Robert; Tschersich, Nikolai (2009): IAB-Haushaltspanel im Niedrigeinkommensbereich Welle 2 (2007/08). Methoden- und Feldbericht. FDZ-Methodenreport 08/2009, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Christoph, Bernhard; Müller, Gerrit; Gebhardt, Daniel; Wenzig, Claudia; Trappmann, Mark; Achatz, Juliane; Tisch, Anita and Gayer, Christine (2008): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Volume I: Introduction and Overview. Wave 1 (2006/2007). FDZ-Datenreport 05/2008 (EN), Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Couper, M. P.; Ofstedal, Mary Beth (2009): Keeping in Contact with Mobile Sample Members. In: Lynn, Peter (Ed.) Methodology of Longitudinal Surveys, Chichester: Wiley.

Couper, Mick P. (1998): Measuring Survey Quality in a CASIC Environment. In: Proceedings of the Survey Research Methods Section, American Statistical Association, p. 41–49.

Dorner, Matthias; Heining, Jörg; Jacobebbinghaus, Peter; Seth, Stefan (2010): The Sample of Integrated Labour Market Biographies. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 130, No. 4, p. 599–608.

Felderer, Barbara; Müller, Gerrit; Kreuter, Frauke; Winter, Joachim (2012): The Effect of Monetary Incentives on Attrition Bias in a Household Panel, institut für Arbeitsmarkt- und Berufsforschung, Nürnberg, unpublished manuskript.

Gebhardt, Daniel; Müller, Gerrit; Bethmann, Arne; Trappmann, Mark; Christoph, Bernhard; Gayer, Christine; Müller, Bettina; Tisch, Anita; Siflinger, Bettina; Kiesl, Hans; Huyer-May, Bernadette; Achatz, Juliane; Wenzig, Claudia; Rudolph, Helmut; Graf, Tobias; Biedermann, Anika (2009): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Volume I: Introduction and Overview. Wave 2 (2007/2008). FDZ-Datenreport 06/2009 (EN), Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Graf, Tobias (2007): Bedarfsgemeinschaften 2005 und 2006: Die Hälfte war zwei Jahre lang durchgehend bedürftig. IAB-Kurzbericht 17/2007, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Groves, Robert M.; Heeringa, Steven G. (2006): Responsive design for household surveys: tools for actively controlling survey errors and costs. In: Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 169, No. 3, p. 439–457.

Groves, Robert M.; McGonagle, Katherine A. (2001): A Theory Guided Interviewing Training Protocol Regarding Survey Participation. In: Journal of Official Statistics, Vol. 17, No. 2, p. 249–265.

Hartmann, Josef; Brink, Kathrin; Jäckle, Robert; Tschersich, Nikolai (2008): IAB-Haushaltspanel im Niedrigeinkommensbereich. Methoden- und Feldbericht. FDZ-Methodenreport 07/2008, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Jesske, Birgit; Quandt, Sylvia (2011): Methodenbericht Panel Arbeitsmarkt und Soziale Sicherung PASS. 4. Erhebungswelle – 2010 (Haupterhebung). FDZ-Methodenreport 08/2011, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Jesske, Birgit; Schulz, Sabine (2013): Methodenbericht Panel Arbeitsmarkt und Soziale Sicherung PASS. 6. Erhebungswelle – 2012. FDZ-Methodenreport 10/2013, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Jesske, Birgit; Schulz, Sabine (2012): Methodenbericht Panel Arbeitsmarkt und Soziale Sicherung PASS. 5. Erhebungswelle – 2011. FDZ-Methodenreport 11/2012, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Kiesl, Hans (2010): Kalibrierte Hochrechnung für das Panel "Arbeitsmarkt und soziale Sicherung", unpublished. Nürnberg.

Kreuter, Frauke (2013): Improving Surveys with Paradata. Analytic Uses of Process Information. New Jersey: Wiley.

Kreuter, Frauke; Müller, Gerrit (forthcoming): A Note on Improving Process Efficiency in Panel Surveys with Paradata. In: Field Methods.

Kreuter, Frauke; Müller, Gerrit; Trappmann, Mark (2010): Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. In: Public Opinion Quarterly, Vol. 74, No. 5, p. 880–906.

Kueppers, Rolf (2005): MOSAIC von microm. In: Grözinger, G.; Matiaske, W. (Eds.) Deutschland Regional. Sozialwissenschaftliche Daten im Forschungsverbund, München: Hampp, p. 95–104.

Laflamme, Francois; Maydan, Mike; Miller, Andrew (2008): Using Paradata to Actively Manage tge Survey Data Collection Process. In: Proceedings of the Survey Research Methods Section, American Statistical Association, p. 630–637.

Laurie, Heather; Smith, Rachel; Scott, Lynne (1999): Strategies for Reducing Nonresponse in a Longitudinal Panel Survey. In: Journal of Official Statistics, Vol. 15, No. 2, p. 269–282.

Lynn, Peter; Kaminska, Olena (2010): Criteria for developing non-response weight adjustments for secondary users of complex longitudinal surveys. Paper presented at the XXI. International Workshop on Household Survey Nonresponse, Nürnberg.

Müller, Gerrit (2011): Wechsel des Erhebungsinstituts und Adressweitergabe mit Widerspruchsverfahren im IAB-Haushaltspanel PASS. In: Methoden, Daten, Analysen. Zeitschrift für Empirische Sozialforschung, Vol. 5, No. 2, p. 207–226.

Peytchev, Andy; Riley, Sarah; Rosen, Jeff; Murphy, Joe; Lindblad, Mark (2010): Reduction of Nonresponse Bias through Case Prioritization. In: Survey Research Methods, Vol. 4, No. 1, p. 21–29.

Rudolph, Helmut; Trappmann, Mark (2007): Design und Stichprobe des Panels „Arbeitsmarkt und Soziale Sicherung“ (PASS). In: Promberger, Markus (Ed.) Neue Daten für die Sozialstaatsforschung: Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung“, Vol. 12/2007 of IAB Forschungsbericht, Nürnberg, p. 60–101.

Schnell, R.; Dietz, C. (2006): CATI-RAT. Multimediale Interviewerschulung für CATI Interviewer. Center for Quantitative Methods and Survey Research, Universität Konstanz, unpublished.

Schnell, Rainer (2007): Alternative Verfahren zur Stichprobengewinnung für ein Haushaltspanelsurvey mit Schwerpunkt im Niedrigeinkommens- und Transferleistungsbezug. In: Promberger, Markus (Ed.) Neue Daten für die Sozialstaatsforschung: Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung", Vol. 12/2007 of IAB Forschungsbericht, Nürnberg, p. 33–59.

Schnell, Rainer; Gramlich, Tobias; Mosthaf, Alexander; Bender, Stefan (2010): Using complete administration data for nonresponse analysis: The PASS survey of low-income households in Germany. Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data.

Schouten, Barry; Cobben, Fannie; Bethlehem, Jelke (2009): Indicators for the representativeness of survey response. In: Survey Methodology, Vol. 35, No. 1, p. 101–113.

Spiess, Martin; Rendtel, Ulrich (2000): Combining an Ongoing Panel with a New Cross-Sectional Sample. Diskussionspapiere / Discussion Papers 198, Deutsches Institut für Wirtschaftsforschung, Berlin.

Särndal, Carl-Erik; Swensson, Bengt; Wretman, Jan (1992): Model Assisted Survey Sampling. New York: Springer.

Stata Corp (2007): Survey Data Reference Manual Release 10. Stata Press, College Station.

Trappmann, Mark; Beste, Jonas; Bethmann, Arne; Müller, Gerrit (2013): The PASS Panel Survey After Six Waves. In: Journal for Labour Market Research, online first.

Trappmann, Mark; Christoph, Bernhard; Achatz, Juliane; Wenzig, Claudia (2007): "Labour Market and Social Security": A New Panel Study for Research on Long Term Unemployment. Paper presented at the International Conference of the German Association of Political Economy.

Trappmann, Mark; Christoph, Bernhard; Achatz, Juliane; Wenzig, Claudia; Müller, Gerrit; Gebhardt, Daniel (2009): Design and stratification of PASS. A New Panel Study for Research on Long Term Unemployment. IAB-Discussion Paper 05/2009, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Trappmann, Mark; Gundert, Stefanie; Wenzig, Claudia; Gebhardt, Daniel (2010): PASS: a household panel survey for research on unemployment and poverty (forthcoming). In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 130, No. 4, p. 609–622.

Wagner, Gert G.; Frick, Joachim R.; Schupp, Jürgen (2007): The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 127, No. 1, p. 139–169.

Watson, Nicole; Wooden, Mark (2009): Identifying Factors Affecting Longitudinal Survey Response. In: Lynn, Peter (Ed.) Methodology of Longitudinal Surveys, Chichester: John Wiley & Sons, p. 157–181.

Imprint

FDZ-Datenreport 07/2013

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Heiner Frank

Technical production
Heiner Frank

Download
http://doku.iab.de/fdz/reporte/2013/DR_07-13.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Arne Bethmann
Institut für Arbeitsmarkt-
und Berufsforschung (IAB)
Regensburger Str. 104
90478 Nürnberg
Tel.: +49 (0) 911/179-2307
mailto:Arne.Bethmann@iab.de