FDZ·Methodenreport

08/2015
EN

Methodological aspects of labour market data

Quick Start File for the Panel "Labour Market and Social Security" (PASS)

Analysing the PASS data using SPSS/PASW

Benjamin Fuchs,
Sarah Lödel,
Michael Otto

Bundesagentur für Arbeit

Quick Start File for the Panel "Labour Market and Social Security" (PASS)
Analysing the PASS data using SPSS/PASW

Benjamin Fuchs (Institut für Arbeitsmarkt- und Berufsforschung)
Sarah Lödel (Friedrich-Alexander-Universität Erlangen-Nürnberg)
Michael Otto (Friedrich-Alexander-Universität Erlangen-Nürnberg)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Summary

This FDZ-Methodenreport is meant to help users of SPSS/PASW handling the data of the Panel Study "Labour Market and Social Security" (PASS). It uses practical examples as can be found in the User Guide of PASS (Bethmann/Fuchs/Wurdack, 2013: Part II), which was written with a focus on STATA users. Therefore, it draws heavily on the work of Gebhardt/Bethmann (2013a), Gebhardt/Bethmann (2013b), Gebhardt/Bethmann (2013c) and Trappmann (2013).

Keywords: PASS, Panel, SPSS, PASW, Quick Start File

1 Key variables

Adapted for SPSS/PASW from Gebhardt, Daniel; Bethmann, Arne (2013a): Key variables. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 68–72.

In the PASS study (Trappmann et al., 2013), as well as in other panel studies, 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. 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.

1.1 Key variables and their connection to the structure of the SUF

The datasets of PASS 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 IDs 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*.[1] 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 *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 original household and form a new split-off household.

Using only the IDs – *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.

---

[1] 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.

Additional information is required to clearly identify an observation, which depends on the format of the specific dataset in question.

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. *alter1* for a persons' age in wave 1, *alter2* 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 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 biography spells contain two observations for a person if this person reported two episodes employment, unemployment or other activities.

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 1 and 2. For further information on their meaning and on how to use them, see the corresponding chapter in Berg et al. (2013). 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.

Table 1: Overview of key variables in the SUF

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 2: Key variables in the datasets of the SUF

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*.

1.2 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 two relevant key variables – the household number (*hnr*) and the wave indicator (*welle*) – must be used.

```
GET FILE='PENDDAT.sav'.
SORT CASES BY hnr (A) welle (A).
MATCH FILES
    /FILE=*
    /TABLE='HHENDDAT.sav'
    /KEEP pnr TO frwunsch hhtyp
    /BY hnr welle.
```

It must be mentioned that the household dataset contains households for which no personal interviews were available. These are re-interviewed households without personal interviews in the respective wave. Due to our specification of the household dataset as the table within the "MATCH FILES" command, these cases need not to be dropped in a next step. This is a main difference to the procedure in STATA, where these cases would have to be dropped within an additional command.

1.3 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.

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr welle.
```

There is a perfect match of the household dataset and the household weights. An inspection of the recently added variables from the household weights shows that there are no system-missing values. For each household that was interviewed in a certain wave an observation from the weighting dataset was merged. See chapter 4 on the use of the weights.

1.4 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.

```
GET FILE='PENDDAT.sav'
    /KEEP=pnr, welle, pintdat.
FORMATS welle (F8.0).
CASESTOVARS
    /ID=pnr
    /INDEX=welle.
VARIABLE LABEL
pintdat.1 'Datum des Personeninterviews in Welle 1'
pintdat.2 'Datum des Personeninterviews in Welle 2'
pintdat.3 'Datum des Personeninterviews in Welle 3'
pintdat.4 'Datum des Personeninterviews in Welle 4'
pintdat.5 'Datum des Personeninterviews in Welle 5'
pintdat.6 'Datum des Personeninterviews in Welle 6'.
EXECUTE.
SAVE OUTFILE='PINTDAT.sav'.
GET FILE='bio_spells.sav'.
SELECT IF spelltyp = 1.
MATCH FILES
```

```
/FILE=*
/TABLE='PINTDAT.sav'
/BY pnr.
```

In the SUF of the 6th wave, 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 and some were not asked about their employment owing to a filter. Again, these cases need not to be dropped in a further step due to our specification of the "MATCH FILES" command.

2 Register data

Adapted for SPSS/PASW from Gebhardt, Daniel; Bethmann, Arne (2013b): Register data. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 73–75.

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 SUF 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 "VARSTOCASES" command in SPSS. 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.[2] 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.

2.1 Household register

All of the households which have been successfully surveyed at least once in the sense of PASS are contained in the household register. Accordingly, households from the gross samples of the individual waves which were not successfully surveyed and households that

---

[2] These are described later in this chapter.

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.

2.2   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. 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.

2.3 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 then stored temporarily and merged with the observations from the first two waves of the household dataset. Both datasets are already sorted according to *hnr*, so there is no need to use the "SORT CASES" command.

```
GET FILE 'hh_register.sav'
/KEEP hnr, alg2samp, hnettok1, hnettok2.
SELECT IF (hnettok1 = 1 & hnettok2 = 1 & alg2samp = 1).
EXECUTE.
DELETE VARIABLES hnettok1, hnettok2.
SAVE OUTFILE '\Datensätze\hh_register_vorb1.sav'.
GET FILE 'HHENDDAT.sav'.
SELECT IF welle = 1 | welle = 2.
MATCH FILES
    /FILE=*
    /TABLE='hh_register_vorb1.sav'
    /BY hnr.
FREQUENCIES alg2samp.
```

An examination of the *alg2samp* variable indicates that 6210 observations 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.

2.4 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.

```
GET FILE= 'hh_register.sav'
    /KEEP hnr pnrzp1 pnrzp2 pnrzp3.
COMPUTE welle1 = 0.
COMPUTE welle2 = 0.
COMPUTE welle3 = 0.
IF (pnrzp1 > 0) welle1 = 1.
IF (pnrzp2 > 0) welle2 = 2.
IF (pnrzp3 > 0) welle3 = 3.
EXECUTE.
VARSTOCASES
    /MAKE pnrzp FROM pnrzp1 pnrzp2 pnrzp3
    /MAKE welle FROM welle1 welle2 welle3.
EXECUTE.
SELECT IF pnrzp > 0.
RENAME VARIABLES (pnrzp = pnr).
COMPUTE hhvorst = 1.
EXECUTE.
SORT CASES BY pnr (A) welle (A).
SAVE OUTFILE='\hh_register_vorb_2.sav'
    /COMPRESSED.
GET FILE= 'PENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hh_register_vorb2.sav'
    /BY pnr welle.
EXECUTE.
RECODE hhvorst (SYSMIS = 0).
EXECUTE.
```

```
FREQUENCIES hhvorst.
```

All of the cases for which the merging was successful were the head of the household in the particular wave and are flagged via the variable *hhvorst*. The cases for whom the merge was not successful are not the heads of households, which is why we recoded the system-missing values to "0" for this variable.

3 Spell data

Adapted for SPSS/PASW from Gebhardt, Daniel; Bethmann, Arne (2013c): Spell data. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 76–80.

In all waves, the SUF of PASS included spell datasets on the household level as well as on the individual level (see table 3 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 at 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

Table 3: Overview of the spell datasets in the SUF

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 Euro ▪ 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)

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 SUF (*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.[3] 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[4] or when the person who answered the household questionnaire in the last interview moved out.[5]

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

---

3   E. g. in wave 2 the employment spells were recorded for the first time in the personal interviews. The respondents were asked to report episodes since January 2005. In wave 3 this date was altered to January 2006.

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

5   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.

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.[6] 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 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.

3.1 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

---

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

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 recorded 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.

```
GET FILE 'alg2_spells.sav'.
COMPUTE hoehebez = -3.
EXECUTE.
DO REPEAT var = AL20800, AL20801, AL20802, AL20803, AL20804, AL20805.
IF (var ~= -3 & var ~= -9) hoehebez = var.
END REPEAT.
EXECUTE.
```

4 Weights

Adapted for SPSS/PASW from Trappmann, Mark (2013): Weights. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 81–99.

4.1 General remarks on the use of the weights

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 the basic version of SPSS, contrary to STATA, *probability weights* cannot be set using the *surveyset* command as shown in Trappmann (2013). 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. In the basic version of SPSS, the user has to use the following command to get valid point estimates:

```
WEIGHT BY wqX.
```

Here *wq*X stands for the adequate weight for the intended analyses. However, this command cannot take into account the impact of the complex sampling desing of PASS on the standard errors. Therefore it enables only correct point estimation but incorrect interval estimation.

4.2 Use of the cross-sectional weights

In this section examples are given on how to use the cross-sectional weights for different purposes. All of the cross-sectional weights in PASS 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 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.

4.2.1 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*. In the example below, 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 "WEIGHT BY" command has to be carried out, and then the projected value can be calculated:

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr welle.
SELECT IF (wqbahh > 0).
WEIGHT BY wqbahh.
TEMPORARY .
SELECT IF (sample = 1 & welle = 1).
FREQUENCIES HLS0800a.
```

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. 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:

```
TEMPORARY.
SELECT IF (sample = 1 & welle = 6).
FREQUENCIES HLS0800a.
```

Approximately 52.4% of the households receiving benefits in July 2006 had a car at the time of the interview in the 6th wave.[7] 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. 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:

```
TEMPORARY.
SELECT IF (alg2abez = 1 & welle = 6 & sample = 1).
FREQUENCIES HLS0800a.
```

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 27.9% have a car. This value has 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.

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

---

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

Unemployment Benefit II a household contains as of the sampling date. [8] The fastest way to do this is to multiply the household weights by this value.

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr welle.
SELECT IF (wqbahh > 0 & nbgbezug > 0).
COMPUTE bgweight = wqbahh*nbgbezug.
EXECUTE.
WEIGHT BY bgweight.
TEMPORARY.
SELECT IF (welle = 1 & sample = 1).
FREQUENCIES HLS0800a.
```

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[9] in households receiving benefits who have a migration background (variable *migration*).

```
GET FILE='PENDDAT.sav'.
```

---

[8]   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.

[9]   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.

```
MATCH FILES
    /FILE=*
    /TABLE='pweights.sav'
    /BY pnr welle.
SELECT IF(wqbap > 0).
WEIGHT BY wqbap.
TEMPORARY.
SELECT IF (welle = 1).
FREQUENCIES migration.
```

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"[10] 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*)

```
TEMPORARY.
SELECT IF (welle = 1 & fb_vers = 1).
FREQUENCIES migration.
```

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:

```
MATCH FILES
    /FILE=*
    /TABLE='p_register.sav'
    /BY pnr.
SELECT IF (pnetto1 = 2 | pnetto1 = 3).
```

---

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

```
TEMPORARY.
SELECT IF (bgbezs1 = 1 & fb_vers = 1 & welle = 1).
FREQUENCIES migration.
```

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%).

4.2.2 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:

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr welle.
WEIGHT BY wqhh.
TEMPORARY.
SELECT IF (welle = 1).
FREQUENCIES HLS0800a.
```

or alternatively with the population sample weights:

```
WEIGHT OFF.
WEIGHT BY wqmihh.
TEMPORARY.
SELECT IF (welle = 1  & wqmihh > 0).
FREQUENCIES HLS0800a.
```

In the first case, the percentage of households with a car is 75.9%, and in the second case 75.6%. The estimation precision is slightly better 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.

4.2.3 Analyses on benefit recipients at different points in time

Section 4.2.1 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 4.2.1 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.

Households in receipt of benefits in July 2011

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr welle.
SELECT IF(wqbahh > 0).
WEIGHT BY wqbahh.
TEMPORARY.
SELECT IF (alg2abez = 1 & welle = 6).
FREQUENCIES HLS0800a.
```

30.5% of all households in receipt of benefits in July 2011 had a car on the interview date of the 6th wave.

Individuals in receipt of benefits in July 2011

```
GET FILE='PENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='pweights.sav'
    /BY pnr welle.
MATCH FILES
    /FILE=*
    /TABLE='p_register.sav'
    /BY pnr.
SELECT IF(wqbap > 0).
WEIGHT BY wqbap.
TEMPORARY.
SELECT IF (bgbezs6 = 1 & fb_vers = 1 & welle = 6).
FREQUENCIES migration.
```

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 (i.e., *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.

```
GET FILE='HHENDDAT.sav'.
MATCH FILES
    /FILE=*
    /TABLE='hweights.sav'
    /BY hnr, welle.
WEIGHT BY wqhh.
TEMPORARY.
SELECT IF alg2abez = 1 & welle = 6.
FREQUENCIES HLS0800a.
```

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

```
WEIGHT OFF.
WEIGHT BY wqbahh.
TEMPORARY.
SELECT IF alg2abez = 1 & welle = 6 & wqbahh > 0.
FREQUENCIES HLS0800a.
```

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. An analysis on car ownership in households receiving Unemployment Benefit II in July 2011, for which we can only work with the BA register data sample, would result in a 95% confidence interval of 28.2% to 32.9%. For the survey date, we would obtain a substantially larger 95% confidence interval of 28.6% to 35.5% (results from an estimation using STATA).

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.[11]

```
GET FILE='p_register.sav'.
AGGREGATE
```

---

[11] 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.

```
        /OUTFILE=MODE=REPLACE
        /BREAK=bgnr6
        / hnr=MEAN(hnr6)
        / bgbezb6=MEAN(bgbezb6).
RECODE bgbezb6 (-5 = 0).
EXECUTE.
SORT CASES BY hnr(A).
AGGREGATE
        /OUTFILE=MODE=REPLACE
        /BREAK= hnr
        / nbgbezak=SUM(bgbezb6).
EXECUTE.
SAVE OUTFILE 'hnr_nbgbezak.sav'.
GET FILE 'HHENDDAT.sav'.
MATCH FILES
        /FILE=*
        /TABLE='hweights.sav'
        /BY hnr, welle.
SELECT IF welle = 6.
MATCH FILES
        /FILE=*
        /TABLE='hnr_nbgbezak.sav'
        /BY hnr.
SELECT IF nbgbezak > 0.
COMPUTE bgw_akt = wqhh*nbgbezak.
EXECUTE.
WEIGHT BY bgw_akt.
TEMPORARY.
SELECT IF alg2abez = 1.
FREQUENCIES HLS0800a.
```

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 respondents. Thus the underreporting of benefit recipiency[12] 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

---

[12] 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.

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 be merged again 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.

```
GET FILE='HHENDDAT.sav'
    /KEEP = hnr, welle, alg2abez.
SAVE OUTFILE 'alg2_info.sav'.
GET FILE='PENDDAT.SAV'.
MATCH FILES
    /FILE=*
    /TABLE='pweights.sav'
    /BY pnr welle.
SORT CASES BY hnr (A) welle (A).
MATCH FILES
    /FILE=*
    /TABLE='alg2_info.sav'
    /BY hnr welle.
SORT CASES BY pnr (A) welle (A).
WEIGHT BY wqp.
TEMPORARY.
SELECT IF alg2abez =1 & welle = 6 & fb_vers = 1.
FREQUENCIES migration.
```

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.

```
SELECT IF welle > 5.
MATCH FILES
    /FILE=*
    /TABLE='p_register.sav'
    /BY pnr.
TEMPORARY.
SELECT IF bgbezb6 = 1 & fb_vers = 1.
FREQUENCIES migration.
```

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. 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.

4.2.4   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 4 provides an overview. The total weights are used in all cases.

4.3   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.

Table 4: 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

4.3.1 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 consists of 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 the weights from the first wave and the re-participation probabilities from wave 1 to wave 2 are stored.

```
GET FILE='pweights.sav'.
SELECT IF welle = 1.
SAVE OUTFILE 'pweights1.sav'.
```

Now the individual dataset is retrieved. We have decided to run the analyses in wide format and therefore we 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.

```
GET FILE='PENDDAT.sav'
    /KEEP=pnr, hnr, welle, PA0300.
FORMATS welle (F8.0).
EXECUTE.
CASESTOVARS
    /ID=pnr
    /INDEX =welle.
EXECUTE.
```

Now the datasets are merged.

```
RENAME VARIABLES (hnr.1 = hnr).
EXECUTE.
SORT CASES BY pnr (A).
MATCH FILES
    /FILE=*
    /TABLE='pweights1.sav'
    /BY pnr.
EXECUTE.
```

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. We only retain cases which have a positive value on the variable *ppbleib*. That is, we only keep those who participated in PASS at least two times and were therefore given a longitudinal weight.

```
COMPUTE rel_zufr = 2.
IF (PA0300.2 = PA0300.1 & PA0300.1 >= 0 & PA0300.2 >= 0) rel_zufr = 1.
IF (PA0300.2 < PA0300.1 & PA0300.1 >= 0 & PA0300.2 >= 0) rel_zufr = 0.
IF (PA0300.1 < 0 | PA0300.2 < 0) rel_zufr= -1.
VALUE LABELS rel_zufr
-1 'in mind. 1 Welle keine Angabe'
0 'W2 weniger zufrieden als W1'
1 'W1 und W2 gleich zufrieden'
2 'W2 zufriedener als W1'.
EXECUTE.
SELECT IF ppbleib > 0.
```

Finally the longitudinal weight is constructed and the weighted analysis follows.

```
COMPUTE wp1_2 = wqp*ppbleib.
EXECUTE.
WEIGHT BY wp1_2.
FREQUENCIES rel_zufr.
WEIGHT OFF.
```

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.

4.3.2 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.

```
COMPUTE wbap1_2 = wqbap*ppbleib.
EXECUTE.
WEIGHT BY wbap1_2.
TEMPORARY .
SELECT IF wqbap > 0.
FREQUENCIES rel_zufr.
WEIGHT OFF.
```

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 that 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.

4.3.3   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.

```
MATCH FILES
    /FILE=*
    /TABLE='p_register.sav'
    /BY pnr.
WEIGHT BY wp1_2.
TEMPORARY.
SELECT IF bgbezb2 = 1 & bgbezb1 = 1.
FREQUENCIES rel_zufr.
```

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.

```
TEMPORARY.
SELECT IF bgbezb2 = 0 & bgbezb1 = 1.
FREQUENCIES rel_zufr.
TEMPORARY.
SELECT IF bgbezb2 = 1 & bgbezb1 = 0.
FREQUENCIES rel_zufr.
```

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.

4.3.4 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.

```
GET FILE='hweights.sav'.
SELECT IF welle = 1.
SAVE OUTFILE 'hweights1.sav'.
GET FILE='HHENDDAT.SAV'
    /KEEP = hnr uhnr welle HLS0800a.
CASESTOVARS
    /ID=hnr
    /INDEX=welle.
MATCH FILES
    /FILE=*
    /TABLE='hweights1.sav'
    /BY hnr.
SELECT IF wqhh > 0.
```

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

```
COMPUTE auto_neu =9.
IF (HLS0800a.1 = 2 & HLS0800a.2 =1) auto_neu = 3.
IF (HLS0800a.1 = 1 & HLS0800a.2 =1) auto_neu = 2.
IF (HLS0800a.1 = 2 & HLS0800a.2 =2) auto_neu = 1.
IF (HLS0800a.1 = 1 & HLS0800a.2 =2) auto_neu = 0.
IF (HLS0800a.1 < 0 | HLS0800a.2 < 0) auto_neu = -1.
MISSING VALUES auto_neu (9).
VALUE LABELS auto_neu
    3 'Auto angeschafft'
    2 'Auto behalten'
    1 'weiterhin kein Auto'
    0 'Auto abgeschafft'
    -1 'in mind. einer Welle keine Angabe'.
EXECUTE.
```

Finally the weight is constructed and the table produced.

```
SELECT IF hpbleib > 0.
COMPUTE wh1_2 = wqhh*hpbleib.
EXECUTE.
WEIGHT BY wh1_2.
FREQUENCIES auto_neu.
```

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 split-off households are then 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%.

4.3.5   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 (Lynn/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 (for wave 6, see Berg et al. (2012)) and did not change much from wave 3. Most variables are taken from the previous (observed) wave and are available to users from the SUF. Additional variables from the paradata of the survey (like number of contact attempts) can be supplied to users upon request.

References

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Bethmann, Arne; Fuchs, Benjamin; Trappmann, Mark; Huber, Martina (2013): Codebook and Documentation of the Panel Study "Labour Market and Social Security" (PASS). Datenreport Wave 6. 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 (2012): 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.

Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (2013): User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6. FDZ-Datenreport 07/2013, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Gebhardt, Daniel; Bethmann, Arne (2013a): Key variables. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 68–72.

Gebhardt, Daniel; Bethmann, Arne (2013b): Register data. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 73–75.

Gebhardt, Daniel; Bethmann, Arne (2013c): Spell data. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 76–80.

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.

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.

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.

Trappmann, Mark (2013): Weights. In: Bethmann, Arne; Fuchs, Benjamin; Wurdack, Anja (Eds.) User Guide "'Panel Study Labour Market and Social Security"' (PASS). Wave 6, FDZ-Datenreport 07/2013, Nürnberg: IAB, p. 81–99.

Trappmann, Mark; Beste, Jonas; Bethmann, Arne; Müller, Gerrit (2013): The PASS Panel Survey After Six Waves. In: Journal for Labour Market Research, Vol. 45, No. 4, p. 275–281.

Imprint

Corresponding author:
Benjamin Fuchs
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nuremberg
Phone: +49-911-179-3479
Email: Benjamin.Fuchs@iab.de