NEPS
National Educational Panel Study

Research Data

Starting Cohort 6: Adults (SC6)
SUF-Version 1.0.0
Data Manual

Thomas Leopold, Marcel Raab, Jan Skopek

SPONSORED BY THE

Federal Ministry
of Education
and Research

Copyrighted Material
University of Bamberg, National Educational Panel Study (NEPS), 96045 Bamberg
http://www.bildungspanel.de
Principal Investigator: Prof. Dr. rer. pol. Dr. h. c. Hans-Peter Blossfeld
Vice Managing Director: Prof. Dr. Sabine Weinert
Executive Director of Research: Dr. Jutta von Maurice
Executive Director of Administration: Dipl. sc. pol. Univ. Dipl.-Betriebswirt (FH) Gerd Bolz
Bamberg, 2011

Data Manual

Starting Cohort 6

Adult Education and Lifelong Learning

NEPS Data Center
Thomas Leopold, Marcel Raab, Jan Skopek

December 22, 2011

Research Data Papers
at the NEPS Data Center, Bamberg

This series presents documentation resources prepared to support the work with data from the National Educational Panel Study (NEPS).

This release of scientific use data from Starting Cohort 6 – "Adult Education and Lifelong Learning" was prepared by the staff of the NEPS Data Center. It represents a major collaborative effort. Most notably, over 150.000 lines of code and almost 900 revisions, in approximately 400 days of work, were produced in the process of data preparation and editing. The contribution of the following staff members of the NEPS is gratefully acknowledged:

Data preparating and editing

Daniel Bela (file integration, cross-sectional data, integration of metadata)
Tobias Koberg (anonymisation, regional data, translation)
Thomas Leopold (testing)
Manuel Munz (coding and classification)
Marcel Raab (testing)
Michael Ruland (editing of longitudinal data)
Sebastian Pink (editing of longitudinal data)
Jan Skopek (editing of longitudinal data)
Knut Wenzig (management and editing of metadata, documentation)

Data manual

Thomas Leopold
Marcel Raab
Jan Skopek

We would like to thank numerous researchers who participated in testing a beta version of these data. Their feedback contributed to a significant improvement of the scientific use file for this main release.

National Educational Panel Study (NEPS)
Data Center
Wilhelmsplatz 3
96047 Bamberg, Germany
Contact: userservice.neps@uni-bamberg.de
Web: https://portal.neps-data.de/de-de/datenzentrum.aspx

Table of Contents

1 Introduction

1.1 About this manual

This manual is intended to assist your work with the data of the NEPS Starting Cohort 6 – Adult Education and Lifelong Learning (SC6 Version 1.0.0). We aim at providing a detailed guide of how to use these data for your research. Therefore, our focus is on practical aspects of data usage such as the dataset structure, key variables, and examples of data retrievals.

This manual is not a comprehensive documentation resource. Please consult our website

https://portal.neps-data.de/de-de/datenzentrum.aspx

for background information on the studies, survey instruments, a structured documentation, and many more resources.

We aim at keeping this manual as short and simple as possible. At several places, we reference supplementary documents presenting additional information that we consider essential for working with our data:

- Codebook (Supplement A)
- How-to guides
  - Merging data (Supplement B)
- Technical reports
  - Weighting (Supplement C)
  - RegioInfas (infas geodaten) (Supplement D)
  - Anonymization Procedures (Supplement E)
  - Data Editing (forthcoming) (Supplement F)

You can download these documents here:

https://portal.neps-data.de/de-de/datenzentrum/forschungsdaten/startkohorteerwachsene.aspx

We welcome feedback from our users that will help us improve the quality of this manual and our data for future releases. Please report any feedback to:

userservice.neps@uni-bamberg.de

1.2 Obtaining the data

There are three simple steps to obtain the data of this release:

- Sign the data use contract and mail it to us (click here for instructions: https://portal.neps-data.de/de-de/datenzentrum/datenzugangswege.aspx)
- After approval, sign in as a registered NEPS user

- Access the data via one of our three access modes (see below)

Depending on which access mode(s) you choose, you will find all further instructions required to access the data on our website.

1.3 Three modes of data access

We offer you three modes of access to the data:

- Download from our website,
- RemoteNEPS (remote access via a virtual desktop),
- and on-site access.

These three solutions are designed to support the full range of users' interests and maximize data utility while complying with strict standards of confidentiality protection.

Sensitive data

Each access mode corresponds to a specific level of data sensitivity. Files that are offered for download include data with the highest level of anynomization. These data are available to registered users from the web portal via a secure connection. Files offered via RemoteNEPS contain more sensitive data within a controlled environment. The analysis of information in high resolution (e.g., fine-grained regional information) is only provided on-site in Bamberg where these data are available within a secure site. For details on the access modes, see our website at

https://portal.neps-data.de/en-us/datacenter/dataaccess.aspx

This concept of data dissemination translates into an "onion-shaped" model of datasets: The most sensitive data ("on-site") that include weakly anonymized information in high resolution represent the outer layer, with "remote access" and "download" levels being subsets of these data. That is, any data contained within a less sensitive level is also included in the higher level(s).

An overview on which types of data are offered at each of these levels as well as detailed information on how the "on-site", "remote access" and "download" versions of the data were generated can be found in Supplement E (see 1.1).

File Format

All files are available in Stata and SPSS format with bilingual variable labels and value labels (German and English). Data stored in Stata format contain both languages within one file (see section 7). SPSS files are delivered separately in both languages.

1.4 Publications with NEPS data

If you publish with NEPS data, it is mandatory to quote the following reference:

Blossfeld, H.-P., H.-G. Roßbach, and J. von Maurice (eds.) (2011). "Education as a Lifelong Process – The German National Educational Panel Study (NEPS)", Zeitschrift für Erziehungswissenschaft: Special Issue 14.

In addition, publications using data from this release must include the following acknowledgement:

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 – Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:1.0.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

A digital object identifier (DOI) uniquely identifies each release of NEPS data. The DOI of this release redirects to a landing page providing basic information on the data:

http://dx.doi.org/10.5157/NEPS:SC6:1.0.0

2 Conventions

2.1 File names

The names of the datasets included in this release were defined by a number of conventions which are displayed in Table 2.

Table 2: Naming conventions of file names

Element	Definition
SC[1-6]	**Indicator of starting cohort** 1 = Infants 2 = Kindergarten 3 = 5th grade students 4 = 9th grade students 5 = First-year undergraduate students 6 = Adults
[filename]	**Filename conventions** Prefix: sp = spell file; p = panel file Keyword/mnemonic: A keyword or mnemonic indicates the content of the corresponding file (e.g., spEmp contains employment spells) Filenames of generated datasets do not have a prefix and always start with a capital letter (e.g., *Biography*)

[D,R,O]	**Confidentiality Level**
	D = Download version
	R = Remote access version
	O = On-site version

[#]-[#]-[#] (_beta)	**Version** First digit: denotes the main release number; the main release number is incremented with every wave of a starting cohort; in starting cohort 6, the main release number 1 comprises respondents from the 2009/10 NEPS sample as well as respondents from the 2007/08 ALWA sample; the latter have already been interviewed twice Second digit: indicates major updates; major updates affect the data structure (e.g., release of imputed datasets); updating your syntax files may be necessary Third digit: indicates minor updates; minor updates affect the content of cells but not the data structure; updating your syntax files is not necessary _beta-suffix: this suffix indicates a preliminary release which allows users to test the data in advance of the main release. The beta version is no longer available after the main release.

2.2 Variable names

The variable naming conventions are aimed at ensuring the consistency of variable names across panel waves. They reflect the panel structure of the NEPS data and allow users to conveniently identify variables across waves.

A variable name consists of up to four elements. The principles of the naming conventions are illustrated by the following example. More detailed information is given in Table 3.

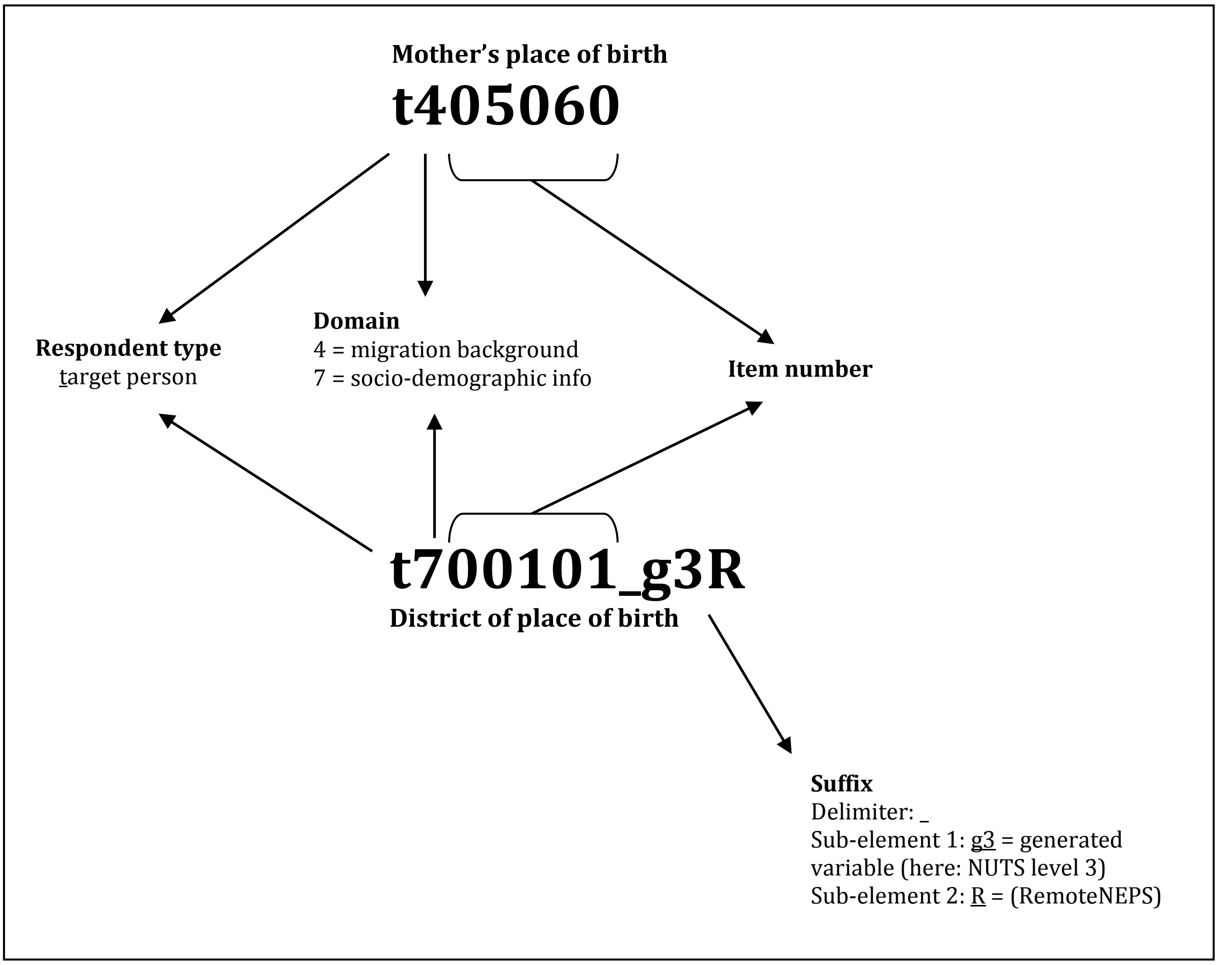

Mother's place of birth
t405060
Respondent type
target person
Domain
4 = migration background
7 = socio-demographic info
Item number
t700101_g3R
District of place of birth
Suffix
Delimiter: _
Sub-element 1: g3 = generated variable (here: NUTS level 3)
Sub-element 2: R = (RemoteNEPS)

Figure 1: Elements of variable names

Table 3: Naming conventions for variable names

Digit	Description
1	Indicates to which **respondent type** the variable refers; in starting cohort 6, this character is always *t* (target person)
2	**Topic/domain** (according to the theoretically coordinated dimensions of the NEPS): 1 = competence development (pillar 1) 2 = learning environments (pillar 2) 3 = educational decisions (pillar 3) 4 = migration background (pillar 4) 5 = returns to education (pillar 5) 6 = working group “interest, self-concept and motivation” 7 = socio-demographic information h = adult education and life-long learning (stage 8) s = basic program; variables developed for the NEPS stages 6 and 8 x = generated variables
3–7	**Item number:** The item number typically consists of four numeric characters plus one alphanumeric character
8–11	**Suffix** (optional): Suffixes are separated from the previous characters by an underscore. There are four types of suffixes: • Version suffixes: Some questions receive minor updates or changes across panel waves. This leads to different versions of similar items. The variable name of the version which will be used in upcoming panel waves – usually the most recent version – does not have a version suffix. The remaining versions are indicated by the following suffixes: o v1 = 2007/08 (ALWA) o v2 = 2009/2010 (Panel = re-interviewed ALWA respondents) o v3 = 2009/10 (First-time respondents) In most cases we were able to integrate earlier versions of variables into the updated version of the variable. If this was not possible, harmonized variables which retain the information common to both versions were generated and marked by the following suffix (see below for examples): o ha = harmonized variable • Suffixes for generated variables: Generated variables are indicated by the suffix _g# (_g1, _g2, etc.). In most cases, the running number after _g is a simple enumerator. However, there are two types of generated variables that assign meanings to these running numbers: regional and occupational variables. o Regional codes based on the Nomenclature of Territorial Units for Statistics (NUTS) ▪ g1 = NUTS level 1 (federal state/Bundesland) ▪ g2 = NUTS level 2 (government region/Regierungsbezirk) ▪ g3 = NUTS level 3 (district/Kreis) o Occupational/prestige codes ▪ g1: KldB 1988 (German Classification of Occupations 1988) ▪ g2: KldB 2010 (German Classification of Occupations 2010)

    - g3: ISCO-88 (International Standard Classification of Occupations 1988)
    - g4: ISCO-08 (International Standard Classification of Occupations 2008)
    - g5: ISEI (International Socio-Economic Index of Occupational Status)
    - g6: SIOPS (Standard International Occupational Prestige Scale)
    - g7: MPS (Magnitude Prestige Scale)
    - g8: EGP (Erikson, Goldthorpe, and Portocarero's class categories)
    - g9: BLK (Blossfeld's Occupational Classification)
    - g10: DKZ 2010 (Documentary Code Number 2010)
    - g11: DKZ 1988 (Documentary Code Number 1988)
    - g12: Coding scheme
    - g13: KKZ (Course code / Kurskennziffer)

- Wide-format suffix:
  Wide-format variables stored in spell files are indicated by the suffix _w# (e.g., _w1, _w2, etc.).
- Confidentiality suffix:
  This suffix pertains to all variables that were anonymized (see 1.4). The suffix indicates a variable's degree of anonymization. This suffix may either stand alone (e.g., country of birth: *t405010_R*) or be combined with other suffixes (e.g., district of place of birth: *t700101_g3R*)
  - O: on site; data on this variable are only available on site
  - R: remote access; data on this variable are available on site or via RemoteNEPS
  - D: download; data on this variable are available via all three modes of access

The following examples illustrate the integration and harmonization of variables that have changed across waves. Table 4 shows that the ALWA study (2007/08) surveyed the current marital status in greater detail than the NEPS in 2009/10. To integrate these two versions, we collapsed categories 1 and 2 of the ALWA (2007/08) item into one category "married". No harmonization was necessary.

Table 4: Panel integration of the variable "current marital status"

t733001	**t733001_v1 (ALWA)**	
1: married	1: married, living together	2: married, separated
2: in a registered partnership		6: in a registered partnership
3: divorced	3: divorced	
4: widowed	4: widowed	
5: single	5: single	

The second example illustrates the harmonization of the variable "mother's place of birth" (Table 5). Here, the NEPS collected more detailed information than the forerunner study ALWA. In the upcoming panel waves, each new respondent will answer the NEPS version of the question (*t405060*). Integrating the ALWA variable (*t405060_v1*) into the NEPS variable was not possible because the response categories varied considerably. Therefore, we generated a harmonized variable (*t405060_ha*) containing the information that both versions have in common.

Table 5: Panel harmonization of the variable "mother's place of birth"

t405060	**t405060_v1 (ALWA)**	**t405060_ha**
1: in Germany (after 1949) 2: in the area that is present-day Germany (before 1950) 3: in Germany's former eastern territories (before 1950)	1: FRG/West Germany 2: GDR/East Germany 3: FRG Germany (after reunification)	1: in Germany
4: abroad (after 1949) 5: in another country (before 1950)	4: abroad	2: abroad

2.3 Missing values

Table 6: Overview of missing codes

Code	**Missing**
Item nonresponse	
–97	refused
–98	don't know
–5/–6/–20	item-specific missing
Not applicable	
–54	not included in sample-specific instrument of this wave
–93	does not apply
.	filtered / system missing
Edition missings (recoded into missing)	
–52	implausible value removed
–53	anonymized
–55	not determinable

We distinguish between three types of missing values (see Table 6):

- *Item nonresponse* occurs if a person did not respond to a question.
  - The most common instances of item nonresponse are refusals (–97) and don't knows (–98).
  - Additional missing codes (–5/–6/–20) pertain to specific nonresponse categories (e.g., –5 "never graduated" for father's school leaving certificate (*t731351*)).
- *Not applicable* denotes missing data that occur because the item does not apply to a person. This category comprises two kinds of missings.
  - The first concerns samples: If a question is not included in a sample-specific questionnaire, the code –54 is assigned to all respondents from this sample.

  - The second concerns individuals: If a question does not apply to a person, it is coded “not applicable” either by the respondent’s or the interviewer’s remark (–93) or automatically by the survey instrument (. = filtered).
- *Edition missings* are defined in the process of data editing.
  - Implausible values are recoded into missing (–52).
  - Sensitive information which is only available via RemoteNEPS and/or on site access is anonymized (–53).
  - Coding schemes are used to generate variables (e.g., occupational coding). If the information from the original data is not sufficient to generate a value, we assign the missing code “not determinable” (–55).

nepsmiss: Recoding missing values in Stata

We offer a Stata ado file on our web portal which automatically recodes all missing values into extended missing values (.a, .b, etc.), and vice versa, while preserving value labels. We generally recommend running *nepsmiss* before any further data preparation. See section 7 for further information.

3 Surveys and sampling

The first release of NEPS SC6 data offers detailed and complete retrospective data on the histories of education, employment, and family. In addition, extensive information was collected on adult education, learning environments, and decision-making processes as well as on subjective well-being and health.

The study population consists of 11,649 individuals born between 1944 and 1986 who were surveyed in 2009/2010. This sample comprises two subsamples:

- ALWA (2007/08) sample: 6,572 respondents from the birth cohorts 1956 to 1986 who were recruited in 2007 by the forerunner study *Working and Learning in a Changing World (ALWA)* conducted by the *Institute for Employment Research (IAB)*, Nuremberg. Note that 77 non-German speaking persons were recruited for the ALWA sample but only interviewed in the second wave (2009/2010).
- NEPS (2009/10) sample: 5,077 respondents recruited for the first wave of the NEPS. Strictly speaking, the NEPS sample consists of two further subsamples:
    - Refreshment sample: drawn from birth cohorts 1956 to 1986
    - Additional sample: drawn from birth cohorts 1944 to 1955

Although this is the initial release of NEPS data, two waves of panel data are already available for 6,495 respondents from the ALWA sample. Table 7 shows which survey instrument was used for which sample in which wave.

Table 7: Overview of samples and survey instruments

	ALWA sample (2007)	**NEPS sample (2009)**
Wave 2007/2008 (ALWA)	ALWA questionnaire *(6,495 respondents)*	
Wave 2009/2010 (NEPS)	NEPS Panel questionnaire *(6,495 respondents)* NEPS questionnaire *(foreign language version) (77 respondents)*	NEPS questionnaire *(5,077 respondents)*

For detailed information on the studies and sampling strategies, see Allmendinger et al. (2011), Antoni et al. (2010), and Aßmann et al. (2011).

Data were collected by computer-assisted personal and telephone interviewing (CAPI & CATI). Figure 2 illustrates the basic structure of the interview.

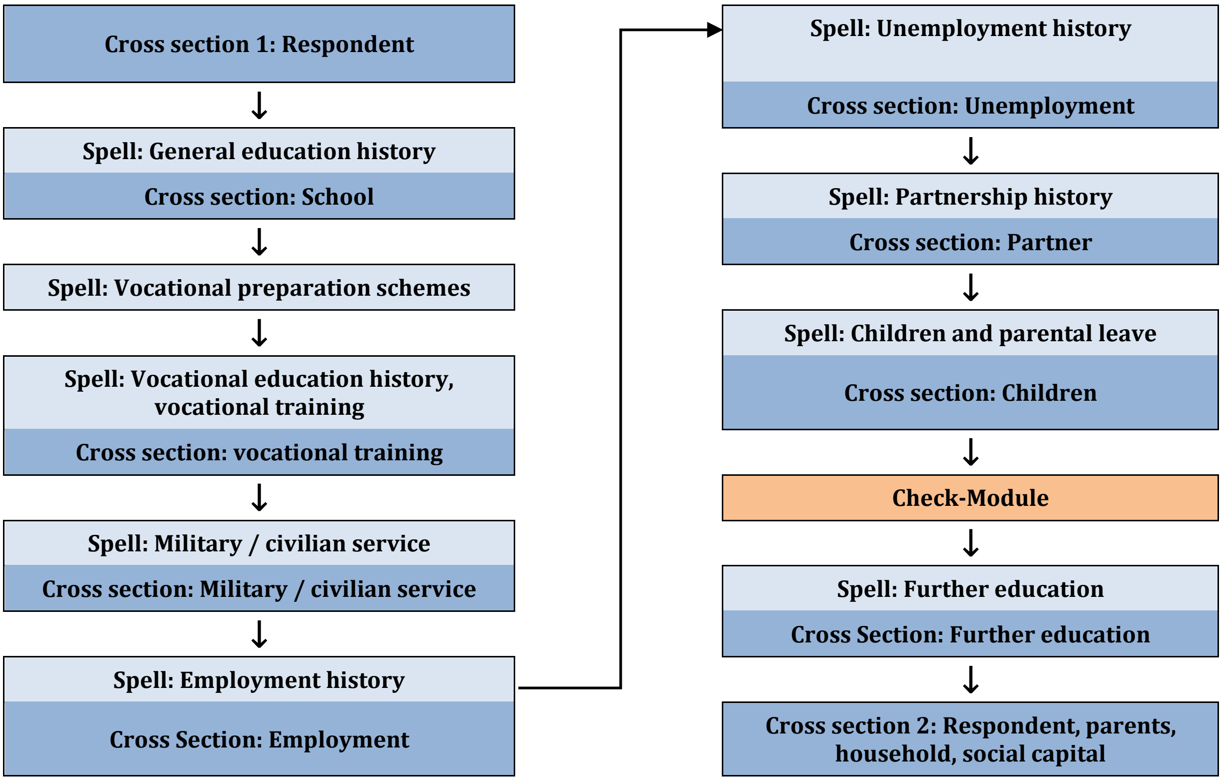

Cross section 1: Respondent
Spell: General education history
Cross section: School
Spell: Vocational preparation schemes
Spell: Vocational education history, vocational training
Cross section: vocational training
Spell: Military / civilian service
Cross section: Military / civilian service
Spell: Employment history
Cross Section: Employment
Spell: Unemployment history
Cross section: Unemployment
Spell: Partnership history
Cross section: Partner
Spell: Children and parental leave
Cross section: Children
Check-Module
Spell: Further education
Cross Section: Further education
Cross section 2: Respondent, parents, household, social capital

Figure 2: Stylized course of interview

The questionnaire began and ended with cross-sectional modules. Between these modules, the questionnaire's main part was devoted to the comprehensive collection of retrospective information on the respondents' life courses. These longitudinal data were collected within ten separate longitudinal modules, most of which were complemented by brief cross-sectional sub-modules. A check module identified and corrected inconsistencies in the sequence of episodes, ensuring the integrity of the life course data.

4 Data structure and datasets

Aims and scope of the NEPS surveys inevitably create complex data. We structured these data in a user-friendly way and generated a number of additional datasets from one or more of the original files to ease the preparation and analysis of life course data. In the following, we distinguish between three types of files according to their data structure: cross-sectional files, panel files, and spell files.

4.1 Cross-sectional files

Basic information: *Basics* (generated file)

This file contains the most recent basic information on each respondent (e.g., socio-demographic variables, current job and household characteristics). These data are generated from the file *pTarget* and a number of spell files (see below). The *Basics* file is updated prospectively. That is, the file is cross-sectional (i.e., one row per person) and always includes updated information from the latest panel wave (if available).

Regional data: *RegioInfas* (generated file)

This file has been generated from the *infas geodaten* database.[1] It comprises geographical information on four regional levels: municipality, postal code, quarters (living areas), and street sections. These data were linked to each respondent by geocoding the sample addresses. Please note that these data are highly sensitive and thus can only be accessed on site. A comprehensive documentation of this dataset is available in Supplement D (see 1.1).

4.2 Panel files

Main panel file: *pTarget*

We merged all cross-sectional information collected at each panel wave into one single dataset *(pTarget)*. This dataset is composed of the two main cross-sectional modules as well as panel data from all cross-sectional sub-modules. These data are stored in long format. That is, one record represents one respondent at one wave.

The file pTarget includes basic socio-demographic information and (repeated) cross-sectional measurements. For example, the first respondent (*ID_t* = 8000215) participated in both waves and therefore has two records, whereas the second respondent (*ID_t* = 8000334) was recruited in the second wave (*wave* = 2) and therefore has only one record.

[1] This database is provided by the infas geodaten GmbH, see: http://www.infas-geodaten.de

```
   ID_t   wave   t700001   t70000y

8000215      1         2      1960
8000215      2         2      1960

8000334      2         2      1959
```

Method dataset: ***Methods***

This dataset offers a variety of information on the data collection (e.g., age, gender and education of the interviewer; interview date; interview duration; incentives), sampling design (e.g., strata variables), and weighting (design weights and calibrated design weights). Detailed information on the calculation and the use of weights is available in Supplement C (see 1.1). You can also find an example in section 0.

4.3 Spell files

Three types of spells: Duration, entity and event spells

This study collected several types of life history data, such as episodes of general education, employment, unemployment, and parental leave. Each of these spell types is stored in a separate dataset. These files always include longitudinal data with each row representing one spell. There are three types of spells:

- *Duration spells* specify a duration spent in a state or episode, such as "employed"
- *Entity spells* pertain to specific entities, such as partners, children, or courses
- *Event spells* defining event times

The variables *spell*, *child, partner,* and *course* are enumerators for each spell of a person:

- *spell* always refers to duration spells
- *child* and *partner* always denote entity spells
- *course* identifies courses in *spFurtherEdu2, spVocTrain, spCourses*, and *spFurtherEdu1* (see examples in section 6 for details); in *spFurtherEdu3* (courses in German), the corresponding identifier is *gcourse*.

Panel updates of spells

Spells are either complete or right-censored. Right-censoring occurs if a spell continues until the time of the interview. The design of the NEPS allows updating right-censored spells prospectively at every panel wave. In this release, such updated spells only concern the ALWA sample. The updating of these spells was executed as follows:

- Spells that were right-censored at the preceding wave (2007/2008 ALWA) were divided into three subspells which are represented by the variable *subspell*:
  - original right-censored episodes from the preceding wave (*subspell* = 1)
  - continued episodes from the panel interview with updated information (*subspell* = 2)
  - harmonized episodes (*subspell* = 0 & *spgen* = 1; see below); in most cases, these edited episodes include the latest information from the panel subspell (*subspell* = 2). However, if this information was either "filtered / system

missing" or "not included in sample-specific instrument of this wave" (–54), these missing values from the panel subspell were replaced by data from the previous subspell (*subspell* = 1). For a few selected variables, there were exceptions to these rules which were guided by plausibility criteria (see below for an example and the forthcoming technical report on data editing for further details).

The main advantage of this procedure is that it retains all information from the original spells while at the same time offering a convenient way of obtaining a harmonized spell data structure. The variable *subspell* is coded 0 both for completed and harmonized spells. Therefore, you can easily obtain a harmonized spell structure by selecting all observations that satisfy the condition

$$\textit{subspell} = 0$$

We generally recommend executing this selection at the start of your data preparation unless you are specifically interested in subspell information. However, be aware that data of harmonized spells may come from different waves because these spells always include the latest valid information available. There is another caveat: Do not use this selection if you work with information stored in wide format (see example below).

The following example illustrates the identification and selection of subspells as well as the logic of harmonized episodes. Employment spells of two persons are displayed. The first respondent (*ID_t* = 8001204) has two employment spells. The second spell was updated prospectively in the panel and was therefore divided into three subspells: *subspell* = 1 represents the right-censored spell from the preceding wave (*wave* = 2007/2008 ALWA); *subspell* = 2 denotes the continued spell from the current wave. The variable monthly net income (*ts23410*) shows the most common rule according to which continued episodes were harmonized. The respondent's income increased from 363 Euros on the first subspell (ALWA) to 400 Euros in the second subspell (NEPS). The harmonized episode (*subspell* = 0) includes the most recent valid information, that is, 400 Euros. The variable occupational status (*ts23204_v1*) is an example for the second rule of harmonizing continued episodes. This item had the value 10 (semi-skilled worker) in the first subspell (ALWA) and was "not included in the sample-specific instrument" (–54) of wave 2 (NEPS). In such cases (i.e., -54 or system missing / filtered), the most recent valid information from the first subspell was retained in the harmonized episode. The second respondent (*ID_t* = 8001138) illustrates an exception to these rules. This person has three employment spells of which the second was updated prospectively. Data on income is available in the first subspell (3800 Euros) and missing (system missing / filtered) in the second subspell. For this variable, it was considered not plausible to replace the missing income from the second subspell by data from the first subspell. Consequently, data on income is missing in the harmonized episode.

```
  +---------------------------------------------------------------------------------------------------------+
  |   ID_t   wave   spell   subspell   spgen   ts2311m   ts2311y   ts2312m   ts2312y   ts23410   ts23204_v1 |
  |---------------------------------------------------------------------------------------------------------|
  | 8001204     1       1          0       0         2      1984         1      1997         .           21 |
  | 8001204     2       2          0       1         3      2006         2      2010       400           10 |
  | 8001204     1       2          1       0         3      2006         1      2008       363           10 |
  | 8001204     2       2          2       0         3      2006         2      2010       400          -54 |
  |---------------------------------------------------------------------------------------------------------|
  | 8001138     1       1          0       0         5      2006         7      2007         .            . |
  | 8001138     2       2          0       1         8      2007         7      2009         .           21 |
  | 8001138     1       2          1       0         8      2007         9      2007      3800           21 |
  | 8001138     2       2          2       0         8      2007         7      2009         .          -54 |
  | 8001138     2       3          0       0         7      2009         2      2010       400          -54 |
  +---------------------------------------------------------------------------------------------------------+
```

Harmonized spells are generated spells and thus can be easily distinguished from complete spells by the indicator variable *spgen*. This variable is coded 1 for all generated spells.

For most analyses it is reasonable to delete the subspells and keep only the generated harmonized episodes (*subspell* = 0 & *spgen* = 1) as well as the complete episodes (*subspell* = 0 & *spgen* = 0). Keeping all observations that satisfy the condition *subspell* = 0 returns a convenient harmonized spell structure with each row representing one episode.

```
  +---------------------------------------------------------------------------------------------------------+
  |   ID_t   wave   spell   subspell   spgen   ts2311m   ts2311y   ts2312m   ts2312y   ts23410   ts23204_v1 |
  |---------------------------------------------------------------------------------------------------------|
  | 8001204     1       1          0       0         2      1984         1      1997         .           21 |
  | 8001204     2       2          0       1         3      2006         2      2010       400           10 |
  |---------------------------------------------------------------------------------------------------------|
  | 8001138     1       1          0       0         5      2006         7      2007         .            . |
  | 8001138     2       2          0       1         8      2007         7      2009         .           21 |
  | 8001138     2       3          0       0         7      2009         2      2010       400          -54 |
  +---------------------------------------------------------------------------------------------------------+
```

Note that some spell datasets include information stored in wide format. Using data from *spChild*, the following example illustrates why harmonized spells (*subspell* = 0) should not be used for these variables. In *spChild*, each record represents one child (entity spells). Coresidence (i.e., sharing a household) episodes of respondents and their children are stored in wide format (denoted by the suffix *_w*).

```
  +-----------------------------------------------------------------------------------------------------------+
  |   ID_t   wave   subspell   child   chcoha_w1   ts3331y_w1   ts3332y_w1   chcoha_w2   ts3331y_w2   ts3332y_w2 |
  |-----------------------------------------------------------------------------------------------------------|
  | 8001929     2          0       1           2         2006         2010           .            .            . |
  | 8001929     1          1       1           1         1987         2005           2         2006         2007 |
  | 8001929     2          2       1           2         2006         2010           .            .            . |
  +-----------------------------------------------------------------------------------------------------------+
```

The variables *chcoha_w1* and *chcoha_w2* count coresidence spells with a child. In this example, the respondent's first interview (*subspell* = 1) yielded information on two coresidence episodes with the first child (*child* = 1) of which the second (*_w2* variables) was right-censored at the interview date. In the second wave, this right-censored spell was updated. Note, however, that information on the updated spell (*subspell* = 2) is now included in the *_w1* variables and information on the second coresidence episode is lost. Therefore, information from the subspells must be retained if variables containing spell information are stored in wide format.

4.3.1 Duration spell files

Integrated life course data: *Biography* (generated file)

The *Biography* file is designed to facilitate the analysis of complex life course data that were collected both retro- and prospectively. This dataset pulls together episodes from the following spell files: *spSchool*, *spVocPrep*, *spMilitary*, *spVocTrain*, *spEmp*, *spUnemp*, *spGap*, and *spParLeave*.

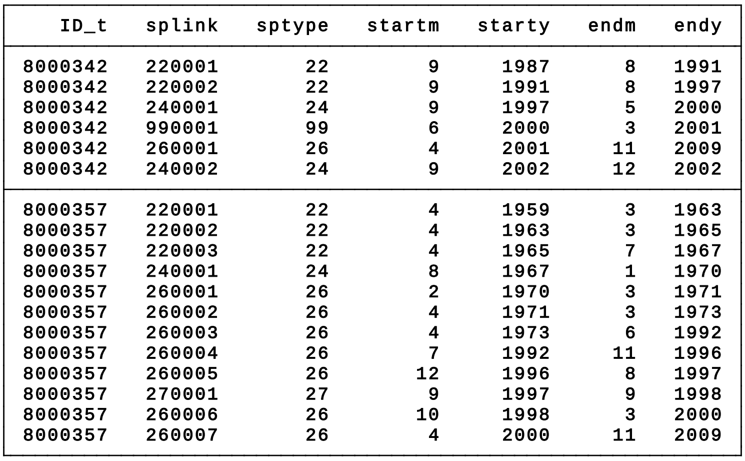
In contrast to the "raw" life course data from each of these modules, the Biography file offers more consistent life course data that are thoroughly checked and edited. During the interview, inconsistencies in individual life course data were identified and corrected by a check module. Further corrections were implemented in the data editing process. Overall, the following measures were taken to ensure the integrity of the life course data in the Biography file:

- All subspells were removed; *Biography* includes only completed, harmonized, or right-censored episodes (i.e., *subspell* = 0, see 4.2).
- Episodes revoked by the respondents during the interview (i.e., disagreement in the introductory question for episode updating in the panel questionnaire) were deleted. Note that the revoked episodes are included in the original spell files and can be identified using the variable *spstat* (91 = spell missing in *Biography*).
- Starting and end dates of episodes were smoothed and corrected:
  - One-month overlaps between adjacent episodes were resolved.
  - Gaps between adjacent episodes which did not exceed two months were closed; gaps of more than two months were defined as specific gap episodes (edition gaps) within the *Biography* file.

The linking variables *ID_t* and *splink* allow matching information from the following duration spell files to the *Biography* file (see section 6 for examples): *spSchool*, *spVocPrep*, *spMilitary*, *spVocTrain*, *spEmp*, *spUnemp*, *spParLeave*, and *spGap*.

Therefore, we recommend using *Biography* as a starting point for life course analyses.

The example displayed below illustrates two respondents' life courses. Episodes follow a clear chronological order: The first respondent (*ID_t* = 8000342) records two school spells (*sptype* = 22) prior to a vocational training episode (*sptype* = 24). There is an edition gap (i.e., a generated spell that bridges a gap in the reported episodes of more than one month; *sptype* = 99) before the first employment spell (*sptype* = 26). This employment spell is right-censored (i.e., it continued until the interview date) and overlaps with a second vocational training episode (e.g., a course) which took place in 2002 (September – December). The second respondent (*ID_t* = 8000357) provided a complete educational and occupational biography without any gaps.

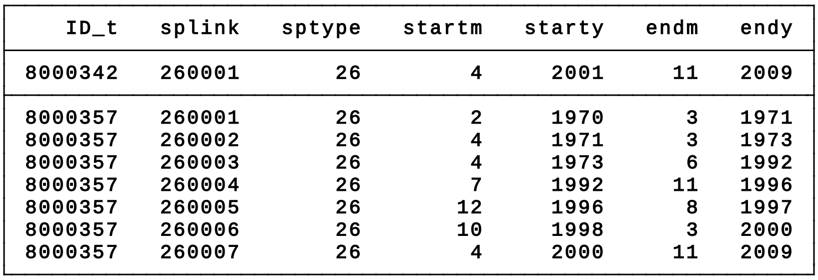


```
     ID_t   splink   sptype   startm   starty   endm   endy

  8000342   220001       22        9     1987      8   1991
  8000342   220002       22        9     1991      8   1997
  8000342   240001       24        9     1997      5   2000
  8000342   990001       99        6     2000      3   2001
  8000342   260001       26        4     2001     11   2009
  8000342   240002       24        9     2002     12   2002

  8000357   220001       22        4     1959      3   1963
  8000357   220002       22        4     1963      3   1965
  8000357   220003       22        4     1965      7   1967
  8000357   240001       24        8     1967      1   1970
  8000357   260001       26        2     1970      3   1971
  8000357   260002       26        4     1971      3   1973
  8000357   260003       26        4     1973      6   1992
  8000357   260004       26        7     1992     11   1996
  8000357   260005       26       12     1996      8   1997
  8000357   270001       27        9     1997      9   1998
  8000357   260006       26       10     1998      3   2000
  8000357   260007       26        4     2000     11   2009
```

Many users may want to restrict their analyses to one life course domain such as the employment career. You can do this by selecting the corresponding spell type. In our example, this spell type is employment (*sptype* = 26).



```
     ID_t   splink   sptype   startm   starty   endm   endy

  8000342   260001       26        4     2001     11   2009

  8000357   260001       26        2     1970      3   1971
  8000357   260002       26        4     1971      3   1973
  8000357   260003       26        4     1973      6   1992
  8000357   260004       26        7     1992     11   1996
  8000357   260005       26       12     1996      8   1997
  8000357   260006       26       10     1998      3   2000
  8000357   260007       26        4     2000     11   2009
```

The next screenshot presents the original dates included in *spEmp*. This allows a comparison between smoothed start and end dates in the *Biography* file (*startm*, *starty, endm, endy*) and the original dates of the complete and harmonized episodes (*subspell* = 0) in *spEmp* (*ts2311m, ts2311y, ts2312m, ts2312y*). Three corrections were executed for the second respondent (*ID_t* = 8000357). Information on the end month of the third employment spell and the start month of the fourth employment spell was not precise in the original data (*ts2312m* and *ts2311m*= 27, “middle of the year”). The upper screenshot shows that this value was replaced by 7 (July) for the start month of the fourth spell. To avoid an overlap, the end date of the previous spell was set to 6 (June). Another overlap in the original data occurred between the end month of the sixth and the starting month of the seventh spell (both *ts2311m* and *ts2312m* have the value 4, “April”). Again, the end date of the previous month was adjusted (*endm* = 3, “March”).

ID_t	subspell	ts2311m	ts2311y	ts2312m	ts2312y
8000342	0	4	2001	11	2009
8000357	0	2	1970	3	1971
8000357	0	4	1971	3	1973
8000357	0	4	1973	27	1992
8000357	0	27	1992	12	1996
8000357	0	12	1996	8	1997
8000357	0	10	1998	4	2000
8000357	0	4	2000	11	2009

General education history: *spSchool*

This module covers each respondent's general education history from school entry until the date of (anticipated) completion, including

- episodes of elementary schooling,
- completed episodes of secondary schooling that led to a school leaving certificate,
- and incomplete episodes of schooling that would have led to a school leaving certificate if they had been completed.

A new episode is generated only if the school type changes. That is, a change from one Gymnasium to another is not recorded. As a result, a single schooling episode may take place at more than one location. In such cases, only information on the last location is included.

A new episode is generated at each school type change even if both schools offer the same certificate. Below you find an example for a person who took four schooling spells to obtain a secondary degree. During the first spell (April 1967 until July 1971), the person was enrolled in elementary school (*ts11204_ha* = 1) which does not award a certificate. Therefore, data on the variables for aspired (*ts11214*) and obtained (*ts11209*) certificates are missing. In the second spell, this person attended a comprehensive school (*ts11204_ha* = 3), aspiring the Abitur (*ts11214* = 5) but not attaining any certificate within this spell (*ts11209* = -5). The third spell represents a (futile) attempt at the Gymnasium (*ts11204_ha* = 5) which lasted from April 1976 until July 1977. Because the school type had changed, a new episode was generated although neither the aspired nor the attainable degree had changed. Back in comprehensive school, the person was finally awarded the Abitur in July 1980. Note that the aspired degree is set to missing because a degree was actually obtained within this episode.

ID_t	spell	ts11204_ha	ts11214	ts11209	ts1111m	ts1111y	ts1112m	ts1112y
8002268	1	1	.	.	4	1967	7	1971
8002268	2	3	5	-5	8	1971	7	1976
8002268	3	5	5	-5	4	1976	7	1977
8002268	4	3	.	5	8	1977	7	1980

Vocational preparation schemes: *spVocPrep*

This module comprises episodes of vocational preparation after general education, including

- pre-training courses,

- basic vocational training years,
- and work preparation courses of the employment agency.

Data were collected on the duration from taking up until completing a vocational preparation scheme, including possible intermissions.

Vocational education history: *spVocTrain*

This module covers all further training, vocational and/or academic, that a respondent ever attended:

- Vocational training and retraining
- Training at technical schools, such as schools of public health, full-time vocational schools (excluding basic vocational training years), other vocational schools, and master craftsmen's colleges.
- Training in specialized fields of medicine
- Accredited training courses to receive licenses
- Conferral of a doctorate or postdoctoral thesis
- Tertiary education at universities, specialized colleges for higher education, colleges of advanced vocational studies, and colleges of advanced administrative and commercial studies.

  Note: Only the main subjects are surveyed. New episodes are generated if

  - a main subject changes over the course of studies
  - the attainable degree changes over the course of studies (e.g., from MA to teaching certification).

  Episodes are continued in case of location changes unless the main subjects change as well.

Training courses for licenses are comparable to courses in the *spCourses*, *spFurtherEdu1*, and *spFurtherEdu2* modules and can therefore be identified by the spell indicator *course*. This enumerator allows linking information about the few courses included in this module to the courses in the modules described below (also see section 6).

Interruptions of vocational training spells, so-called vocational interruption episodes, are stored in wide format. These sub episodes are not yet edited and should therefore be treated with caution.

Courses: *spCourses*

This module comprises courses and trainings attended within the past 12 months during episodes of employment, unemployment, parental leave, military or civilian service as well as episodes from the *spGap* module. The starting and end dates of the spells in this module represent the original episodes (from *spEmp*, *spUnemp*, etc.) in which a course was taken. For each of these episodes, information on up to three courses is included in wide format (variables with the suffix *_w* – see 2.2); *spCourses* comprises all spells from the past 12 months which were recorded in the modules mentioned above. Spells may also be included if no course was taken during this episode. The only criterion for inclusion in the module is that a person provided information on at least one course.

The following example illustrates the data structure of this module. Data from two persons are displayed. The first person (*ID_t* = 8000523) contributes two employment spells to the module. In the past 12 months, s/he attended courses within both spells, amounting to a total of five courses taken. The second person (*ID_t* = 8000645) is represented by one employment spell and one unemployment spell. Note that both spells are included although this person only attended courses during the unemployment spell.

```
     ID_t   splink   subspell   sptype   t27800a   t27800b   t27800c   t27800d   course_w1   course_w2   course_w3

  8000523   260007          0      Emp         4      1990        11      2009           1           2           3
  8000523   260008          0      Emp         4      2001        11      2009           4           5           .

  8000645   260008          0      Emp         9      1999         9      2009           .           .           .
  8000645   270001          0    UnEmp        10      2009        11      2009           1           2           3
```

Note that in *spCourses*, the course enumerator is stored in wide format (*course_w1*, *course_w2*, and *course_w3*), whereas in the other course modules (*spFurtherEdu1* and *spFurtherEdu2*) there is only a single enumerator (*course*) (also see the examples in section 6). Furthermore, it is important to note that the variable *subspell* refers to the original episodes (from *spEmp*, *spUnemp*, etc.). Therefore, you should not use the condition *subspell* = 0 (see 4.3) if you work with this module.

Additional courses: *spFurtherEdu1*

This module contains information on courses attended in addition to courses reported in *spCourses* or in *spVocTrain*. These include both professional trainings (similar to those from *spCourses*) and courses attended for private purposes (e.g., cookery course, yoga course, fortune telling, NLP coaching) within the past 12 months. In contrast to *spCourses*, this module's spells refer to the actual starting and end dates of the courses.

German courses: *spFurtherEdu3*

Information on courses in German as a foreign language is collected only for immigrants. These data are stored in this module. The course enumerator for the German courses is *gcourse*.

Integrated course file: *FurtherEducation* (generated file)

Information about the respondents' participation in further education is distributed across several spell files. The generated file *FurtherEducation* integrates data on all courses from *spCourses* and *spFurtherEdu1* as well as vocational courses and trainings from *spVocTrain* into one consolidated format. In *FurtherEducation*, these courses are stored as duration spells in long format. Start and end dates of courses were imputed if this information was not precise (e.g., "spring") or missing. Data on the content of courses are available as open answers and in a coded version using a classification of the *Federal Employment Agency* (Kompetenzkatalog der Bundesagentur für Arbeit).

All respondents who reported on at least one participation in further education are included in *FurtherEducation*. Note that in contrast to *spCourses* and *spFurtherEdu1*, this file contains not only course participations from the last year but also from the previous life course. The latter are vocational trainings reported in *spVocTrain* that can be classified as courses and trainings related to further education. The variable *course* (course number) allows tracking courses back to the original files *spCourses*, *spFurtherEdu1*, and *spVocTrain*. For a subset of courses that have a course number, further information from *spFurtherEdu2* can be added. Furthermore, there is a second subset of courses that can be linked to spells from *spVocTrain* or *spEmp* because they have been reported in the context of these spells or (in case of spells from *spVocTrain*) are directly derived from them. The variables *ID_t*, *course*, and *splink* allow matching these original spell data to *FurtherEducation*. Table 8 provides an overview which courses are included in *FurtherEducation* and to which spells they can be linked in the original files.

Table 8: Overview of courses included in FurtherEducation

course	**splink**	**Description**
valid	missing	Further education spell reported in the further education module (stored in *spFurtherEdu1*); spell is right-censored or ended within the past 12 months.
missing	24#### (Vocational Training)	Vocational training spells related to further education and participation(stored in *spVocTrain*); spell ended more than 12 months ago.
valid	24#### (Vocational Training)	Vocational training spells of the type “further education and participation” (stored in *spVocTrain*); spell is right-censored or ended within the past 12 months
valid	25#### (Military/Civilian Service)	Further education reported in the course module; triggered by spells in *spMilitary* (courses are stored in *spCourses*); triggering spell is right-censored or ended within the past 12 months
valid	26#### (Employment)	Further education reported in the course module; triggered by spells in *spEmp* (courses are stored in *spCourses*); triggering spell is right-censored or ended within the past 12 months
valid	27#### (Unemployment)	Further education reported in the course module; triggered by spells in *spUnemp* (courses are stored in *spCourses*); triggering spell is right-censored or ended within the past 12 months
valid	29#### (Parental Leave)	Further education reported in the course module; triggered by spells in *spParLeave* (courses are stored in *spCourses*); triggering spell is right-censored or ended within the past 12 months
valid	30#### (Gap)	Further education reported in the course module; triggered by spells in *spGap* (courses are stored in *spCourses*); triggering spell is right-censored or ended within the past 12 months

The following example illustrates the structure of this dataset. In the course of the interview, the respondent reported on a total of three courses. Each of these courses was recorded by a different module and stored in a different dataset. The variable *tx28200* identifies the source dataset of a spell (24 = *spVocTrain*; 31 = *spFurtherEdu1*; 35 = *spCourses*). Finally, the variable *tx28202_g1* includes coded data on the content of courses based on the classification of the *Federal Employment Agency* (Kompetenzkatalog der Bundesagentur für Arbeit).

```
   ID_t   number   course   splink   tx28200   tx2821y   tx2822y               tx28202_g1

8000410        1        .   240002        24      2006      2006               K 1400-025
8000410        2        1   260004        35      2008      2009   -55 (not determinable)
8000410        3        2        .        31      2009      2009             K 090204-005
```

Military / civilian service and voluntary gap years: *spMilitary*

This module includes episodes of military or civilian service as well as gap years taken to do voluntary work in the social or environmental sector. Regular or professional soldiers are considered employed and are therefore included in the employment module.

Employment history: *spEmp*

This extensive module covers all spells of regular employment, including traineeships. Information on second jobs is only collected for activities that continue to the interview date. Vacation jobs, volunteering, and internships are not included.

New episodes are created at the following events:

- Change of employer
- Change of occupation
- Interruption of employment (e.g., unemployment or military service)

Unemployment history: *spUnemp*

This module includes all episodes of unemployment irrespective of whether a person was registered as unemployed or not. Questions on registration of unemployment and receipt of benefits refer both to the beginning and to the end of an unemployment spell.

Parental leave: *spParLeave*

For each child in *spChild* (except for deceased children), information is collected on whether the respondent took a parental leave. Each parental leave episode contributes one record to *spParLeave*.

Parental leaves do not include maternity protection. These periods are added to the corresponding employment episode. As a result, an employment spell is not interrupted if the mother only takes the maternity leave without an additional parental leave.

Gaps: *spGap*

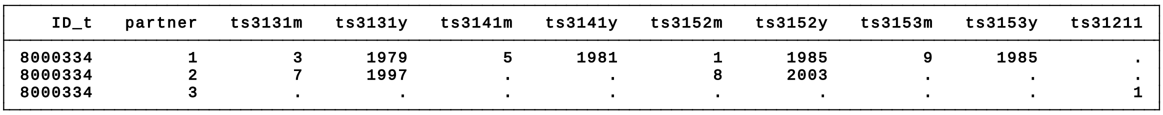
Gaps in individual life courses are identified by a check module. Such gap episodes are included in the *spGap* module. The spells in this file refer to different types of gaps which can be distinguished by the variable *ts29101*. The most common gap episodes are (extended) holidays and looking after home or family.

4.3.2 Entity spell files

History of partners: *spPartner*

This module covers the partnership history of the respondent. Respondents' subjective reports define whether they live in a relationship and whether they cohabit or not. A comprehensive set of additional questions refers to the present partner. For earlier partners, only information on the year of birth and education is available. Information on the current partner is collected regardless of the cohabitation status, whereas previous partners are only included if they cohabitated with the respondent. The enumerator variable *partner* identifies partners "within" respondents. This variable is coded 1 for the first partner and counts upward until the last (current) partner.

The following example illustrates the data structure of this module. The respondent reported on three partners. Partners 1 and 2 are previous partners. Remember that previous partners are only included in this module if they cohabited with the respondent. Cohabitation with partner 1 lasted from May 1979 until January 1985. The respondent married this partner in May 1981 and divorced in September 1985. The second relationship was a consensual union which began in July 1997 and ended in August 2003. No data on cohabitation is available for the third partner. That is, information on this partner is recorded although s/he never cohabited with the respondent. This is only possible for current partners who are included regardless of the cohabitation status. Compared to previous partners, information on current partners is available in greater detail. This example shows one of these variables (*ts31211*) which indicates whether the partner is German or not. This information was not collected retrospectively for previous partners. As a result, data on this variable is missing for partners 1 and 2.



```
   ID_t   partner   ts3131m   ts3131y   ts3141m   ts3141y   ts3152m   ts3152y   ts3153m   ts3153y   ts31211

8000334         1         3      1979         5      1981         1      1985         9      1985         .
8000334         2         7      1997         .         .         8      2003         .         .         .
8000334         3         .         .         .         .         .         .         .         .         1
```

Children: *spChild*

This module contains information on

- all biological, foster, and adopted children of the respondent,
- and every other child that currently lives or has ever lived together with the respondent (e.g., children of former and current partners).

In cases of twins and higher orders of multiple births, separate episodes are generated for each child. Episodes generally refer to the periods in which the respondent and the child shared a household. Because *spChild* is an entity spell file, data on starting and end dates of cohabitation are stored in wide format. The enumerator variable *child* identifies children within respondents. Note that a child episode was skipped in the interview if the respondent reported that the child was deceased.

Take a look at the example below. The first respondent reported on one biological child (*ts33204* = 1). Note that data on the duration of coresidence are stored in wide format. The first coresidence episode lasted from November 1981 until July 2003. Apparently, this child is a “boomerang kid” who returned to the parental home in June 2007 and stayed there until the interview date (right-censored). For the second person, information on a total of six children is included. The variable *ts33204* indicates that only three of these are biological children. Child numbers 4, 5, and 6 are children from one (or more) partner(s) (*ts33204* = .) who are included in the module because they cohabited with the respondent.

ID_t	child	ts33204	ts3331m_w1	ts3331y_w1	ts3332m_w1	ts3332y_w1	ts3331m_w2	ts3331y_w2	ts3332m_w2	ts3332y_w2
8000381	1	1	11	1981	7	2003	6	2007	11	2009
8008539	1	1	5	1980	10	1999	.	.	.	.
8008539	2	1	7	1982	8	1993	.	.	.	.
8008539	3	1	3	1986	11	2008	.	.	.	.
8008539	4	.	5	1992	7	1996	.	.	.	.
8008539	5	.	5	1992	10	2002	.	.	.	.
8008539	6	.	5	1992	10	2006	.	.	.	.

History of children: *Children* (generated file)

This entity spell file was generated from the *spChild* module, offering basic information (e.g., current cohabitation state, cohabitation history) about all biological, step, foster, and adopted children as well as other children with whom the respondent has ever cohabited.

Selected courses: *spFurtherEdu2*

The survey instrument randomly selected two courses from the *spVocTrain*, *spCourses* and *spFurtherEdu1* modules, collecting additional information on these courses (e.g., costs, motivation, certificates). These data are included in *spFurtherEdu2*.

4.3.3 Event spell files

In contrast to duration spells, event spells do not specify the time spent in a certain state (e.g., unemployed) but the point in time at which a transition between two states (i.e., an event) occurred.

Transitions in educational careers: *Education* (generated file)

This generated file provides longitudinal information on transitions in respondents' educational careers. It contains only persons who have an educational degree at a lower secondary level or higher. We used all information on educational attainment from *spSchool* (lower, intermediate, and upper secondary school degrees – Hauptschule, Realschule, (Fach-)Abitur), *spVocPrep* (participation in vocational preparation schemes), and *spVocTrain* (all successfully completed trainings). Three measures of educational attainment are available: CASMIN (variable *tx28101*), ISCED-97 (*tx28103*), and years of education (*tx28102*; derived from CASMIN).

This information is stored as event data in long format. That is, each row represents a transition in at least one classification (CASMIN and/or ISCED-97). Variables on month and year of the transition (*datem* and *datey*) specify the event time. We consider only upward educational transitions in CASMIN levels and upward as well as lateral transitions in ISCED-97 levels (CASMIN is ordinal, whereas ISCED-97 has some nominal elements). Because ISCED-97 and CASMIN follow different concepts, some educational transitions (approximately 7 % in these data) are captured by only one of these classifications.

The variable *sptype* indicates the dataset from which an event (i.e., an educational transition) was generated. For example, information on the transition to Abitur was taken from *spSchool* (*sptype* = 22). You can easily merge data from these original spells to *Education* using the variable *splink*. Note that the ISCED-classification is a little more fine-grained in wave 2 (NEPS) compared to wave 1 (ALWA) because the measures of educational attainment were more differentiated in the NEPS survey instrument.

The following example illustrates the structure of this dataset. The first respondent (*ID_t* = 8000507) obtained a lower secondary degree (Hauptschulabschluss) in March 1966. This degree is represented by the value 1 both in the CASMIN (*tx28101*) and ISCED-97 (*tx28103*) classification. The variable *tx28109* indicates that a change was recorded in both classifications (denoted by the value 3). This always applies to the first event spell of a respondent in this dataset. In September 1969 (second event spell), the respondent completes a vocational training (Lehre). Consequently, CASMIN is set to the value 2 and ISCED-97 is set to the value 4. Because this upward transition concerns both classifications, the variable *tx28109* is again 3. Three years later (September 1972), the respondent experiences a vocational upward transition (e.g., master's qualification, Meister/in). This transition is only captured by the ISCED-97 classification (*tx28103* increases from 4 to 8) but not by the CASMIN classification which remains at the value 2 (i.e., CASMIN does not differentiate between basic and advanced vocational trainings). As a result, *tx28109* is set to the value 2, indicating that only ISCED-97 changed its value. The reverse is true for the fourth (and final) event spell of this respondent in which an educational upward transition is recorded. This change is only captured by the CASMIN classification. The corresponding value of CASMIN (*tx28101*) is 6, indicating that the respondent has attained an A-levels qualification (or equivalent) in addition to the vocational training that had already been completed. Therefore, *tx28109* has the value 1, denoting a change only in CASMIN. Note that the variable sptype specifies the source of the information from which these event spells were generated (22 = *spSchool*; 24 = *spVocTrain*).

```
+---------------------------------------------------------------------------+
|    ID_t   splink   datey   datem   tx28101   tx28103   tx28109   sptype |
|---------------------------------------------------------------------------|
| 8000507   220001    1966       3         1         1         3       22 |
| 8000507   240001    1969       9         2         4         3       24 |
| 8000507   240002    1972       9         2         8         2       24 |
| 8000507   220002    1974       9         6         8         1       22 |
|---------------------------------------------------------------------------|
| 8000512   220001    1968       8         1         1         3       22 |
| 8000512   240002    1974       9         2         4         3       24 |
| 8000512   240004    1986       8         7         9         3       24 |
+---------------------------------------------------------------------------+
```

The following tables show how the ISCED-97 and CASMIN classes are composed.

Table 9: Coding of ISCED-97

Codes in tx28103	ISCED-97		
	Key	**Englisch**	**German**
0	0A/1A	Inadequatly completed general education	kein Abschluss
1	2B	Lower general education	Haupt-, Volksschulabschluss, Berufsvorbereitende Maßnahme
2	2A	Intermediate general education	Mittlere Reife, Realschulabschluss
3	3A	Full maturity certificates (e.g., the Abitur, A-levels)	Fachhochschulreife, Hochschulreife
4	3B	Basic vocational training, Vocational full time school, Health sector school (less than two years), civil servant of the lower grade, vocational basic skills	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse
5	3C	Civil servants of the medium grade	Beamter mittlerer Dienst
6	4A	Full maturity certificates (e.g., the Abitur, A-levels) (second cycle)	Fachhochschulreife, Hochschulreife (second cycle)
7	4B	Basic vocational training, Vocational full time school, Health sector school (less than two years), civil servant of the lower grade, vocational basic skills (second cycle)	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse (second cycle)
8	5B	Diploma (vocational and other specialised academies, College of public administration), Qualification of a two or three year Health-Sector School, Master's/technician's qualification	Fach- und Berufsakademische Abschluss, Verwaltungsfachhochschule, Fachschule des Gesundheitswesens (mindestens zwei Jahre), Meister/Techniker, anderer Fachschulabschluss, Beamter gehobener Dienst
9	5A	Bachelor, Master, Diploma, state examination, civil servants of the highest grade	Bachelor, Master, Diplom, Magister, Staatsexamen, Beamter höherer Dienst
10	6	Doctoral degree and postdoctoral lecture qualification	Promotion

Table 10: Coding of CASMIN

Codes in tx28101	CASMIN		
	Key	**Englisch**	**German**
0	1a	Inadequatly completed general education	Kein Abschluss
1	1b	General elementary education	Hauptschulabschluss ohne berufliche Ausbildung
2	1c	Basic vocational training above and beyond compulsory schooling	Hauptschulabschluss mit beruflicher Ausbildung
3	2b	Intermediate general education	Mittlere Reife ohne berufliche Ausbildung
4	2a	Intermediate vocational qualification, or secondary programmes in which general intermediate schooling is combined by vocational training	Mittlere Reife mit beruflicher Ausbildung
5	2c_gen	General maturity: Full maturity certificates (e.g., the Abitur, A-levels)	Hochschulreife ohne berufliche Ausbildung
6	2c_voc	Vocational maturity: Full maturity certificates including vocationally specific schooling or training	Hochschulreife mit beruflicher Ausbildung
7	3a	Lower tertiary education: Lower level tertiary degrees, generally of shorter duration and with a vocational orientation	Fachhochschulabschluss
8	3b	Higher tertiary education: The completion of a traditional, academically orientated university education	Universitätsabschluss

Marital history: *MartialStates* (generated file)

This file was generated from the *spPartner* module. It contains event spell data on each respondent's marital states. The variable *marstate* distinguishes between three categories: married, divorced, and widowed. Only persons who have married are included in this file.

4.4 Overview of all datasets

Table 11 presents an overview of all datasets included in this release. Note that the number of respondents contained in each file varies markedly because specific modules only apply to certain subgroups of respondents ("universes"). For example, *spPartner* only includes those who currently have a partner and/or have ever cohabited with a partner.

Table 11: Overview of all datasets

File	Content	Type	Universe	N (persons)
Basics	Most recent basic information on the respondent	Cross-sectional	All respondents	11,649
RegioInfas	Regional information	Cross-sectional	All respondents	11,649
pTarget	Socio-demographic information Repeated cross-sectional measurements (panel data)	Panel	All respondents	11,649
Methods	Data collection, sampling design, weights	Panel	All respondents	11,649
Biography	Integrated life course data	Duration spell	Resp. with at least one spell in *spSchool, spVocPrep, spMilitary, spVocTrain, spEmp, spUnemp, spParLeave*, and/or *spGap*	11,646
spSchool	General education history	Duration spell	Resp. who attended general school and/or received a general school certificate	11,637
spVocPrep	Vocational preparation schemes	Duration spell	Resp. who attended vocational preparation schemes	954
spVocTrain	Vocational education history	Duration spell	Resp. who (at least started) vocational training	11,220
spCourses	Courses and trainings	Duration spell	Resp. who attended training courses during employment, unemployment, parental leaves, military/civilian service, or gap episodes.	4,923
spFurtherEdu1	Additional courses	Duration spell	Resp. who attended further courses	2,232
spFurtherEdu3	Courses in German	Duration spell	Resp. with migration background who ever attended a German course	243
FurtherEducation	Integrated course file	Duration spell	Resp. who reported on at least one participation in further education	6,482
spMilitary	Military / civilian service	Duration spell	Resp. who served in military or civilian service or completed voluntary work in the social or environmental sector	3,820
spEmp	Employment history	Duration spell	Resp. who reported on at least one employment or traineeship	11,518
spUnemp	Unemployment history	Duration spell	Resp. who were unemployed (registered or unregistered) at least once	6,632
spParLeave	History of parental leaves	Duration spell	See *spChild*	3,366
spGap	Gap episodes	Duration spell	Resp. who reported gaps between labour market and educational activities	6,824
spPartner	History of partners in the household	Entity spell	Resp. who currently have a partner or ever cohabitated with a partner	10,752
spChild	History of children in the household	Entity spell	Resp. who have children and/or ever cohabitated with children	8,538
Children	Birth biography	Entity spell	Resp. who have children and/or ever cohabitated with children	8,538
spFurtherEdu2	Detailed information on two randomly selected courses	Entity spell	Additional information on courses in *spVocTrain, spCourses*, and *spFurtherEdu1*	4,475
Education	Longitudinal data on transitions in educational careers	Event spell	Resp. who have an educational degree at a lower secondary level or higher	11,505
MaritalStates	Marital biography	Event spell	All respondents who ever married	8,682

Note

The NEPS invested a lot to ensure the integrity of these data. However, we strongly recommend you to examine the data critically when you work with this release. Furthermore, you should always consult the questionnaire/s to obtain a precise understanding of how the data have been collected. Finally, it is important to note that each additional dataset that we created for the users' convenience was generated on the basis of certain assumptions (e.g., event times in the file *Education* were calculated from the smoothed duration times in the file *Biography*). Please keep these assumptions in mind if you use data from the generated files.

4.5 Merging the data

A number of identifiers allow combining information from different datasets. A unique and never-changing *ID_t* (identifier of target person) is assigned to each respondent. This identifier is required for all matching procedures. In *pTarget*, the variable *wave* further indicates in which wave(s) a respondent was observed. In case of spell data, additional variables are needed to uniquely identify observations within a dataset. There are five basic matching procedures:

1. Use *ID_t* to match data from *Basics* to all other datasets.
2. Use *ID_t* and *wave* to match data from *pTarget* and *Methods* to all other datasets.
3. Use *ID_t* and *splink* to match data from all duration spells (*spSchool, spVocPrep, spMilitary, spVocTrain, spEmp, spUnemp, spCourses, spGap, Education*) to the *Biography* file.
4. Use *ID_t* and *child* to match data from *spParLeave* to data from *spChild.*
5. Use *ID_t* and *course* to match data from *spFurtherEdu2* to *spVocTrain*, *spFurtherEdu1, spCourses* and/or the generated file *FurtherEducation*.

See section 6 for examples on each of these matching procedures. A comprehensive overview of all matching procedures is available in Supplement B (see 1.1):

5 Generated variables and weights

5.1 Coding

All string variables on occupations of respondents, their parents, and partners were coded. Table 12 presents an overview of these coded variables and the variables that are derived from them as well as the CASMIN educational classification which is particularly useful if you are interested in cross-national comparisons.

Table 12: Overview of coded variables

Classification	**Included in**	**Description**
KldB88	*spEmp; spVocTrain; pTarget; spPartner*	German Classification of Occupations 1988 (4-digit)
KldB2010	*spEmp; spVocTrain; pTarget; spPartner*	German Classification of Occupations 2010 (5-digit)
ISCO-88	*spEmp; spVocTrain; pTarget; spPartner*	International Standard Classification of Occupations 1988 (4-digit)
ISCO-08	*spEmp; spVocTrain; pTarget; spPartner*	International Standard Classification of Occupations 2008(4-digit)
BLK	*spEmp; spVocTrain; pTarget; spPartner*	Occupational classification by Blossfeld
ISEI	*spEmp; spVocTrain; pTarget; spPartner*	Metric scale to measure prestige of occupations
SIOPS	*spEmp; spVocTrain; pTarget; spPartner*	Metric scale to measure prestige of occupations
MPS	*spEmp; spVocTrain; pTarget; spPartner*	Magnitude prestige score of occupations (Wegener)
EGP	*spEmp; spVocTrain; pTarget; spPartner*	Class scheme which assigns occupations to classes
CASMIN	*pTarget; spPartner; Basics; Education*	Classification representing differentiated educational attainment and vocational training degrees
ISCED-97	*pTarget; spPartner; Basics; Education*	Classification representing differentiated educational attainment and vocational training degrees
Years of education	*pTarget; spPartner; Basics; Education*	Years of education based on the CASMIN classification

5.2 Weights

Information on the construction of weights and how to use them can be found in the technical report on weighting (Supplement C) and the examples section 0. Note that weights are only available for wave 2 (NEPS).

6 Examples

This section gives some examples of how to merge different datasets from this release (using Stata).

Merging *Basics* with other datasets

Variables from the cross-sectional file *Basics* can easily be merged to all other datasets of this release. In the example shown below we merge data on the respondent's gender and the father's EGP class (when the respondent was aged 15) to the employment spell file (*spEmp*).

```
*Merge information from Basics to other files

/*
Procedure
1. Open spEmp
2. Merge variables from Basics to spEmp with a m:1-merge
*/

***

use "SC6_spEmp_D_1-0-0.dta", clear

merge m:1 ID_t using "SC6_Basics_D_1-0-0", keepusing(t700001 t731453_g8) keep(1 3)

tab _merge // gender and father's EGP class were matched to all 59,266 spells
```

Merging *pTarget* with other datasets

Virtually everyone who works with spell files will draw on information stored in *pTarget*, such as the respondents' gender. If you merge a spell file with *pTarget*, you should keep in mind that *pTarget* is a long-format file. If you want to merge time-constant information such as gender with a spell file (e.g., *spEmp*), you only need information from one record of each respondent in *pTarget*.

```
/*
Procedure
1. Open pTarget and select only last wave record for each respondent using "duplicates
drop"
2. Save a temporary version (helpfile) of the reduced pTarget
3 Open spEmp and add the gender variable from our helpfile using a m:1-merge
*/

***

use "SC6_pTarget_D_1-0-0_beta.dta", clear
keep ID_t wave t700001  // only keep the required variables
gsort +ID_t -wave // sort by ID_t ascending and wave descending
duplicates drop ID_t, force // drops all but the first record of each ID_t that has > 1
observations in the data

***

tempfile helpfile // defines the local macro "helpfile" as a temporary file
save `helpfile', replace // saves the information to be merged

***

use "SC6_spEmp_D_1-0-0_beta.dta", clear
merge m:1 ID_t using `helpfile', keep(1 3)

tab _merge // gender was merged to all 59,266 spells
```

Note that merging time-variant panel variables (e.g., income) to a spell file is much more complicated because you have to deal with different time axes in the files you intend to merge. Whereas a row in *pTarget* represents a year in which the respondent participated in the survey, a record in a spell file corresponds to one specific episode (e.g., an employment spell) or entity (e.g., a partner).

Merging duration spells with *Biography*

This example illustrates how to merge the smoothed and corrected starting and end dates of the *Biography* file with the employment history (*spEmp*). Because the *Biography* file includes only harmonized or completed episodes, you have to delete subspells (*subspell* = 1 and *subspell* = 2) before merging data from duration spells with the *Biography* file.

```
/*
Procedure 1: Merge Biography to spEmp
1. Open spEmp and delete subspells 1 and 2
2  Merge spEmp with Biography using ID_t and splink as key variables (1:1-merge)
*/

***

use "SC6_spEmp_D_1-0-0_beta.dta", clear
keep if subspell == 0 // only keep harmonized and completed spells

***

merge 1:1 ID_t splink using "SC6_Biography_D_1-0-0_beta", keep(1 3)

tab _merge // 47,368 episodes merged; 146 episodes are included only in master file

************************

/*
Procedure 2: Merge spEmp to Biography
1. Open Biography and select employment spells
2  Merge spEmp to Biography using ID_t and splink as key variables (1:m-merge)
*/

***

use "SC6_Biography_D_1-0-0_beta", clear
keep if e_sptype == 26 // keep employment spells

***

*Note: 1:m-merge is necessary because spEmp contains subspells 1 and 2
merge 1:m ID_t splink using "SC6_spEmp_D_1-0-0_beta.dta", keep(1 3)

keep if subspell == 0 // keep harmonized and completed spells

tab _merge // 47,368 episodes merged
```

The example illustrated two different approaches to merging data from the *Biography* file with the *spEmp* module. Note that the first approach yields more observations after merging has been completed. This is because *spEmp* still contains 146 episodes revoked by the respondents during the interview.

Merge *spParLeave* with *spChild*

If you want to link information about the respondents' children to the corresponding parental leave episodes, you have to use the key variables *ID_t* and *child*. In this example, we merge information on the child's gender (*ts33203*) and year of birth (*ts3320y*) to the parental leave file.

```
/*
Procedure
1. Open spChild and select spells that are completed and harmonized
2. Save a temporary version (helpfile) of the reduced spChild
3. Open spParLeave and add information from the helpfile using a m:1-merge
*/

***

use "SC6_spChild_D_1-0-0", clear
keep if subspell == 0 // only keep harmonized and completed spells
keep ID_t child ts3320y ts33203 // only keep the required variables

***

tempfile helpfile // defines the local macro helpfile as a temporary file
save `helpfile', replace // saves the information to be merged

***

use "SC6_spParLeave_D_1-0-0", clear
merge m:1 ID_t child using `helpfile' , keep(1 3)

tab _merge // information on 6,290 children merged
```

Merge courses from *spVocTrain*, *spFurtherEdu1*, or *spCourses* with *spFurtherEdu2*

Data on courses are stored in several files of this release. Some basic information on courses which the respondent attended during the 12 months before the interview can be found in *spVocTrain*, *FurtherEdu1*, and *spCourses*; *spFurtherEdu2* contains more detailed information on two randomly selected courses from these three files.

If you want to merge *spFurtherEdu2* to the other modules, remember that courses are stored in different formats across the files. In *spVocTrain* and *spFurtherEdu1*, courses are stored in spell format. Therefore, they can be easily merged with *spFurtherEdu2* using *ID_t* and course as key variables (see examples 1 and 2). However, courses in *spCourses* are stored in wide format. Here the data must be reshaped into long format before they can be merged with *spFurtherEdu2* (see example 3).

```
*Example 1: spFurtherEdu2 to spVocTrain

/*
Procedure
1. Open spVocTrain and select the course spells
2. Add detailed information on the courses (spFurtherEdu2) where possible
*/

***

use "SC6_spVocTrain_D_1-0-0", clear
keep if !missing(course)

***

merge 1:1 ID_t course using "SC6_spFurtherEdu2_D_1-0-0_beta", keep(1 3)

tab _merge // details available for 25 courses
```

```
*Example 2: spFurtherEdu2 to spFurtherEdu1

/*
Procedure:
  Open spFutherEdu1 and add detailed information
  on the courses (spFurtherEdu2) where possible
*/

use "SC6_spFurtherEdu1_D_1-0-0", clear

merge 1:1 ID_t course using "SC6_spFurtherEdu2_D_1-0-0_beta", keep(1 3)

tab _merge // details available for 1,906 courses
```

```
*Example 3: spFurtherEdu2 to reshaped spCourses

/*
Procedure
1. Open spCourses and select course-specific variables
2. Reshape dataset from wide to long format
3. Prepare the reshaped dataset for merging
4. Add detailed information on the courses (spFurtherEdu2) where possible
*/

***

use "SC6_spCourses_D_1-0-0", clear

drop  wave t278000-t271001    // drop unimportant variables
drop t272011_w2R t272011_g1w2 t272011_w3R t272011_g1w3 t272011_w1R t272011_g1w1

***

reshape long course_w  t271011_w t271012_w t271013_w, i(ID_t splink subspell)
drop if missing(course_w)     // drop generated rows which don't store any course
information
drop _j

***

*Removing the _w-suffixes

*Alternative 1: use regular expression

foreach var of varlist *_w {
        local newvar = regexr("`var'","(_w[O]?)$","")
        rename `var' `newvar'
}

*Alternative 2: simply use the ado "renvars" (net install dm88_1.pkg):
*renvars *_w, postdrop(2)

***

merge 1:1 ID_t course using "SC6_spFurtherEdu2_D_1-0-0", keep(1 3)

tab _merge // details available for 5,051 courses
```

Accounting for sample stratification and using weights

The file *Methods* contains variables for sample stratification as well as weights. This information can be used to correctly estimate population parameters.

```
*Example: Accounting for sample stratification and using weights

/*
Procedure:
  1) Prepare Methods file to obtain sampling and weighting information
  2) Merge this information to the Basics file
*/

use "SC6_Methods_D_1-0-0.dta", clear
drop if wave==1 // remove wave records from 2007/2008 (ALWA)
keep ID_t psu stratum  weight_design_std    // keep relevant variables
tempfile weights
save `weights', replace

use "SC6_Basics_D_1-0-0.dta", clear
merge 1:1 ID_t using `weights', assert(3) nogen

nepsmiss _all

* do some descriptive analyses using standardized design weights
tab t700001 [aweight=weight_design_std]
* define complex survey data structure to adjust standard errors
svyset psu [pweight=weight_design_std], strata(stratum) singleunit(certainty)
// estimate the mean and standard error of age at interview
svy: mean tx29000
// regress net household income on education, gender, and civil state
svy: regress t510010_g1 i.t700001 i.tx28101 i.tx27000
```

7 Tools for Stata users

Our Stata files offer variable labels and value labels both in German and in English. You can easily switch between these languages using the `label language` command.

```
label language en
label language de
```

Furthermore, we have developed two Stata programs (ado files) to ease work with our data. You can obtain these ado files from our repository using the following command:

```
net install nepstools, from(http://neps-data.de/stata)
```

nepsmiss: Recoding missing values

This program automatically recodes and labels all missing values into extended missing values (.a, .b, etc.). In this example, we run nepsmiss on the variable *t731454*, decoding all negative values (-54, -97, -98) into Stata's extended missings (.c, .b, .a).

```
nepsmiss t731454
```

ID_t	wave	t731454
8010851	2	-97
8012254	1	-54
8002388	2	-98
8012254	2	5
8002388	1	1

ID_t	wave	t731454
8010851	2	.b
8012254	1	.c
8002388	2	.a
8012254	2	5
8002388	1	1

We generally recommend running *nepsmiss* on all variables (`nepsmiss _all`) before any further data preparation.

infoquery: Display survey questions

This program displays the survey question that corresponds to a variable in a dataset. Note that infoquery will produce no output for some derived variables.

```
infoquery t405060
```

```
query result for variable t405060:

t405060[questiontext_de]:
Wo ist Ihre Mutter (Stiefmutter / diese Person) geboren?

t405060[questiontext_en]:
Where was your mother (stepmother/ this person) born?
```

8 Further information

Please visit our web portal for further information and comprehensive documentation resources such as CATI questionnaires, how-to guides, technical reports, and the codebook.

https://portal.neps-data.de/de-de/datenzentrum/forschungsdaten/startkohorteerwachsene.aspx

For further support, please contact the NEPS data center:

Web:

https://portal.neps-data.de/de-de/datenzentrum/kontaktdatenzentrum.aspx

E- Mail:

userservice.neps@uni-bamberg.de

Phone:

+49-(0)951-863-3511 (Mo-Fr 10:00-12:00 and 14:00-16:00)

Participation in the NEPS user trainings

Furthermore, the NEPS data center offers training courses on a regular basis. These courses introduce the research design of the NEPS, the structure of datasets, terms and conditions of data usage, issues of privacy and data protection, and so on. A central module of the courses consists of hands-on work with the NEPS data supervised by our staff. As skill levels, research interests, and methods vary greatly across users and disciplines, we will offer a comprehensive portfolio of seminars ranging from introductory topics on a rather general level to advanced methodological courses.

References

Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R., & Ruland, M. (2011). Adult education and lifelong learning. In *H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice* (Eds.), Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS) (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2010). Arbeiten und Lernen im Wandel. Teil I: Überblick über die Studie (FDZ Methodenreport No. 05/2010). Nürnberg.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S. & Blossfeld, H.P. (2011). Sampling designs of the National Educational Panel Study: challenges and solutions. In *H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice* (Eds.), Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS) (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. Wiesbaden: VS Verlag für Sozialwissenschaften.