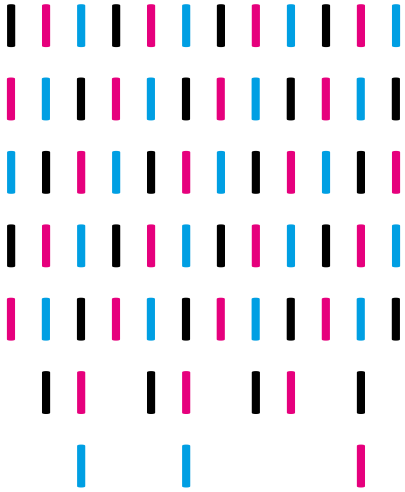

NEPS *SURVEY PAPERS*

Steffen Schindler

CONSTRUCTION OF AN INFORMATION-BY-AGE DATASET BASED ON STARTING COHORT 6 OF THE NATIONAL EDUCATIONAL PANEL STUDY CONTAINING EDUCATION AND EMPLOYMENT TRAJECTORIES ACROSS THE LIFE COURSE

NEPS *Survey Paper* No. 120
Bamberg, June 2025

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS *Survey Paper* series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS *Survey Papers* are edited by a review board consisting of the scientific management of LIfBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section “Publications“) and at www.lifbi.de/publications.

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

Construction of an Information-by-Age Dataset based on Starting Cohort 6 of the National Educational Panel Study containing Education and Employment Trajectories across the Life Course

Steffen Schindler, University of Bamberg

E-mail address:

steffen.schindler@uni-bamberg.de

Bibliographic data:

Schindler, S. (2025). *Construction of an Information-by-Age Dataset based on Starting Cohort 6 of the National Educational Panel Study containing Education and Employment Trajectories across the Life Course* (NEPS Survey Paper No. 120). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP120:1.0

Construction of an Information-by-Age Dataset based on Starting Cohort 6 of the National Educational Panel Study containing Education and Employment Trajectories across the Life Course

Abstract

This paper provides a complementary description of a Stata syntax package for data transformation of the Starting Cohort 6 (SC6) of the German National Educational Panel Study (NEPS). The syntax addresses two major aspects. First, as the NEPS SC6 has not been systematically edited and cleaned by the data producers, the syntax provides extensive cleaning procedures with respect to life-course information on educational participation and attainment, labor market participation and earnings. Second, while the data cleaning is accomplished in the long format, the syntax also provides code to transfer the data into an information-by-age format (wide format). This means that – for every respondent – age-specific variables provide information on educational attainment or labor market outcomes at a given age. For some further operations (e.g. imputation), storing longitudinal information in this wide format can be preferable. In addition, this format might be more intuitive for users who are not familiar with the handling of spell data or data in long format.

Keywords

education, labor market information, starting cohort 6, data cleaning, age-specific information

1. Introduction

This paper provides a complementary description of a Stata syntax package for data transformation of the Starting Cohort 6 (SC6) of the German National Educational Panel Study (NEPS).[1] The syntax has been developed as part of the research project "The Role of Educational Upgrading in the Formation of Social Inequality"[2] with the ultimate goal of transferring the data into a format that provides a basis for multiple imputation of longitudinal information on educational and occupational trajectories and on earnings. As the SC6 is highly complex data that requires much effort to get it into shape for analyses, a secondary goal in the development of the syntax was to make it openly available for other NEPS users who share similar research interests.[3] The syntax addresses two major aspects. First, as the NEPS SC6 has not been systematically edited and cleaned by the data producers, the syntax provides extensive cleaning procedures with respect to life-course information on educational participation and attainment, labor market participation and earnings. Second, while the data cleaning is accomplished in the long format, the syntax also provides code to transfer the data into an information-by-age format (wide format). This means that – for every respondent – age-specific variables provide information on educational attainment or labor market outcomes at a given age. For some further operations (e.g. imputation), storing longitudinal information in this wide format can be preferable. In addition, this format might be more intuitive for users who are not familiar with the handling of spell data or data in long format.

The syntax files of the package have become quite extensive. For that reason, this paper is meant as a complementary guide with some further explanations. The syntax itself is written in a simple way to facilitate comprehension of the rationales behind the code. Note that data cleaning oftentimes requires decisions to recode the data in one way or another. These decisions have been documented through comments in the syntax files, but it is, of course, up to the users to recode the data differently. The code has been extensively checked and double-checked, however – due to its complexity – it is very likely that it still contains errors. If users spot such errors, feedback will be highly appreciated. The syntax has been written with Stata version 16.1 and is based on version 14.0.0 of the scientific use file of NEPS SC6. The following pages will provide a guide through the different syntax files and follow the order from the Master-do-file.

[1] This paper uses data from the National Educational Panel Study (NEPS, see Blossfeld & Rossbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network.

[2] The project received funding from Deutsche Forschungsgemeinschaft (DFG) - Project number 442414757.

[3] The Stata do-files can be downloaded from the NEPS Survey Paper website or zenodo.org: https://doi.org/10.5281/zenodo.15227908 or www.uni-bamberg.de/sozbalv/forschung/syntax/.

2. Master-do-file

`MASTER.do` provides the starting point of the syntax package. It starts all relevant single do-files and runs them in a sequential order, since several do-files build on each other. It is divided into different sections, which will also structure the remainder of this paper.

To run the syntax, users need to define globals for the folders where they store data, do-files and log files and for the NEPS SC6 version.

`global data`	specify the path to the folder where you store the original NEPS data files
`global dofile`	specify the path to the folder where you store the package of do-files
`global savedta`	specify the folder where you want to save data that will be generated by the syntax
`global savelog`	specify the folder where you want to save log files
`global SUF`	specify the version of the SC6 scientific use file in the format “14-0-0” (this corresponds to the number at the end of the original NEPS data files)

3. Download ado files

The syntax requires that users have installed several user-written commands. The Master-do-file comprises the download and installation of these ado-files.

`nepstools`	this package has been developed by the NEPS research data center and provides a number of specific commands that facilitate the handling of NEPS data

4. Store relevant NEPS datasets with NEPSmiss

One of the ado-files specifically developed for NEPS is the command `nepsmiss`. In the original NEPS scientific use files, missing values are represented through negative numerical values. The `nepsmiss` command transfers these values into the Stata convention for missing values (.a, .b, etc.). The do-file `nepsmiss.do` applies this procedure to all relevant NEPS SC6 data sets and stores them in the `$savedta` folder.[4]

5. Recode and store basic information in wide format

The final data set will contain several variables with time-constant information, such as the date of birth, gender or social origin. Some of these information will be necessary for the recoding of the longitudinal information. For that reason, the recode of time-constant information will be conducted first. All data sets containing time-constant information will be stored in the wide format.

[4] The command `plnepsmiss` by Bittmann (2024) provides a computationally faster way of recoding the missing values but is not implemented in the NEPStools package yet. It can be installed through `net install plnepsmiss,from(https://raw.github.com/fbittmann/plnepsmiss/stable) replace`

5.1 wavepart.do

The do-file `wavepart.do` utilizes information from the original *CohortProfile* dataset and recodes indicators about respondents' participation in the panel study. The data will be stored as `wavepart.dta`. It contains the following variables:

`part_w*`	the binary variables `part_w1` – `part_w14` indicate whether respondents participated in waves 1 to 14
`date_w*`	the variables `date_w1` – `date_w14` contain the date when the interview of the respective wave was conducted (in months counting from January 1960)
`partfirst`	indicates the wave of the first participation, which can be either wave 1 (ALWA sample), wave 2 (1st NEPS main survey), or wave 4 (refreshment).
`partlast`	indicates the wave of the last recorded participation
`partnum`	indicates the number of waves the respondent has participated
`lastint`	indicates the date of the last interview (in months counting from January 1960)[5]

5.2 basicinfo.do

The do-file `basicinfo.do` draws on the original `Basics` dataset and provides information about time-constant properties of the respondents, such as date of birth, sex, social origin or migration history. It will be stored as `basicinfo.dta`. It contains the following variables:

`birthyear`	indicator of respondents' year of birth
`birthdate`	respondents' date of birth (in months since January 1960)
`gender`	indicator of the respondent's gender
`casmin_*`	indicator of parents education based on the CASMIN classification. The subscript _f denotes father's CASMIN, the subscript _m denotes mother's CASMIN, and the subscript _p denotes the highest parental CASMIN
`educyears_*`	indicator of parents education based on the years spent in the education system. The subscript _f denotes father's years of education, the subscript _m denotes mother's years of education, and the subscript _p denotes the highest number of parental years of education
`egp_*`	indicator of parents social class based on the EGP class schema (11-category version). The subscript _f denotes father's EGP, the subscript _m denotes mother's EGP, and the subscript _p denotes the highest parental EGP
`isei_*`	indicator of parents social status based on the International Socio-Economic Index of Occupational Status (ISEI, ISCO-08 version). The subscript _f

[5] All variables that are coded as months counting from January 1960 are stored in the %tm format, so that the actual years and months are displayed.

	denotes father's ISEI, the subscript _m denotes mother's ISEI, and the subscript _p denotes the highest parental ISEI
siops_*	indicator of parents social prestige based on the Standard International Occupational Prestige Scale (SIOPS, ISCO-08 version). The subscript _f denotes father's SIOPS, the subscript _m denotes mother's SIOPS, and the subscript _p denotes the highest parental SIOPS
migration	binary indicator whether respondents have a migration history
migration_gen	indicator of the migration generation based the NEPS migration generation concept
migration_age	indicator of the age of immigration (in years) for first and first-and-a-half generation immigrants

6. Expand spell data to monthly information

The do-file month_episodes.do recodes information from the original Biography dataset. This original dataset provides information about episodes in different states (e.g. school attendance, vocational training, employment etc.) together with their start and end dates. The do-file recodes this spell information into monthly information. This means that, for each respondent, the resulting dataset month_episodes.dta contains a line for each month under observation and a variable indicating the state during that month. When respondents are in multiple states in the same month, the new dataset contains multiple lines per months and a variable indicating the number of the state within that month.

Table 1

Example of Variables Stored in month_episodes.dta

ID_t	month	monthnr	splink	dur	sptype	spms
8000278	1965m9	1	220001	119	School	Main spell
8000278	1965m10	1	220001	119	School	Main spell
8000278	1965m11	1	220001	119	School	Main spell
8000278	1965m12	1	220001	119	School	Main spell
8000278	1966m1	1	220001	119	School	Main spell
...						
8000278	1992m2	1	260005	67	Emp	Main spell
8000278	1992m2	2	240002	9	VocTrain	Side spell
8000278	1992m3	1	260005	67	Emp	Main spell
8000278	1992m3	2	240002	9	VocTrain	Side spell
8000278	1992m4	1	260005	67	Emp	Main spell
8000278	1992m4	2	240002	9	VocTrain	Side spell

Table 1 shows an extract from the dataset with an example of a respondent that has multiple states during the same months (here: employment and vocational training). The dataset contains the following variables:

month	indicator of the month counting from January 1960.

monthnr	indicator of the number of the state, which is larger than one if a respondent has multiple states in the same month
splink	indicator identifying the spell (original spell number)
dur	indicator of the total duration of the spell in months
sptype	indicator of the type of spell
spms	indicator whether is a the main spell or a side spell

7. Recode of information on education

The next step of the recoding procedure is devoted to information on educational trajectories. The recodes of school trajectories and school-leaving degrees on the one hand and vocational training trajectories and vocational degrees (including higher education) will be conducted in two separate steps, following the data structure of the NEPS. Both steps follow the same principles: a first do-file conducts basic recodes to generate variables with the relevant information, a second do-file undertakes plausibility checks and data edits to clean the data from inconsistent information, and a third do-file conducts recodes based on the edited data to provide variables with the requested information for the final information-by-age dataset. In a final step, information on school-leaving and vocational degrees will be combined in one dataset.

7.1 Information on school-leaving degrees

7.1.1 school_basics.do

The do-file school_basics.do provides a first recode of relevant school-related information. It starts from the month_episodes dataset and adds information from the original spSchool dataset, which contains detailed information about school spells. In addition, it draws on information about date of birth and the age of migration from the basicinfo dataset (see above) to help determine whether schools-leaving degrees have been obtained in Germany or abroad. It creates the following variables, which are stored in the dataset school_basics.dta:

agem	age in months
agey	age in years
schooltype	indicator of the type of school attended
schoolcont	binary indicator whether the school episode still continues
schooldegree	indicator of the school-leaving degree obtained at the end of the episode, including categories for: still attending this school, no qualification intended and no qualification obtained
recog	indicator whether the foreign degree has been recognized in Germany
schoolgpa	indicator of grade point average (GPA) of the school-leaving degree on the German grading scale (1 very good – 6 insufficient)

abifield*	indicators of the fields in which the Abitur exams were taken (up to five different fields are possible, depending on the federal state and year)
gerschool	binary indicator whether the school has been attended in Germany
fedschool	indicator of the federal state, in which the school has been attended, including categories for: Germany (not further specified) and missing categories for: indeterminate, abroad

7.1.2 school_plausi.do

The do-file school_plausi.do conducts comprehensive plausibility checks and data cleaning procedures to correct inconsistencies in the school_basics dataset. The first step is an inspection of parallel school spells. There are 240 cases with two parallel spells and 5 cases with three parallel spells in the data (version 14-0-0). *Table 2* provides an example of a case with two parallel school spells as indicated by the column monthnr.

Table 2

Example of a Case with Parallel School Spells

ID_t	splink	schooltype	schooldegree	month	monthnr	dur	agey
8000424	220002	Gymnasium	Abitur	1976m7	1	225	6.5
8000424	220002	Gymnasium	Abitur	1976m8	1	225	6.583333
8000424	220002	Gymnasium	Abitur	1976m9	2	225	6.666667
8000424	220001	Element/Volks	indeterminate	1976m9	1	47	6.666667
8000424	220002	Gymnasium	Abitur	1976m10	2	225	6.75
8000424	220001	Element/Volks	indeterminate	1976m10	1	47	6.75
8000424	220002	Gymnasium	Abitur	1976m11	2	225	6.833333
8000424	220001	Element/Volks	indeterminate	1976m11	1	47	6.833333

In this example, the episodes in elementary school and Gymnasium appear to overlap, which is a rather unrealistic scenario. Considering the young age (displayed in the last column of the table) suggests an input error related to the timing of the episode in the Gymnasium, which typically succeeds the episode in elementary school. In this case, the timing of the Gymnasium episode has been edited, so that it directly follows the episode in elementary school. Similar corrections have been undertaken for the other 244 instances with parallel school spells after case-wise inspection.

The second step in the do-file concerns issues with episodes in elementary school. A first issue relates to respondents that appear to have obtained implausible school-leaving degrees at the end of the elementary school spell. Typically, the elementary school does not award any school-leaving degree in Germany. Hence, in these cases, the correct code in the variable schooldegree should be “no qualification”. A complication arises for respondents that attended the school type *Volksschule*, which comprised 8 school years and combined elementary and lower secondary education. Typically, this school type awarded the *Volksschulabschluss*, which can be considered equivalent to the *Hauptschulabschluss*. If the duration in elementary school plausibly approximated these 8 school years and the degree type was *Hauptschulabschluss*, these cases have not been edited. For any other school-leaving degrees (*Mittlere Reife, Fachhochschulreife, Abitur*), the data have been edited – either by

correcting the variable indicating the school-leaving degree or the variable indicating the school type. A second issue concerns 9,846 cases in the data with indeterminate information about the school-leaving degree upon completion of the episode in elementary school. Taking into account information about the duration of the episode, the regulations in the respective year and the location (Federal Republic of Germany or German Democratic Republic), these cases were coded as “no qualification” or as “Hauptschulabschluss”, whenever one of these categories could be plausible assumed to apply. A third issue relates to cases with multiple spells in elementary school. If any subsequent spell in elementary school indicated that no qualification has been pursued, the same code has been transferred to the previous spell in elementary schools if there was any record of a school-leaving degree.

The third step in the do-file concerns school forms of secondary education with implausible school-leaving degrees. There are 94 cases with Fachhochschulreife or Abitur obtained at lower-secondary schools and 34 cases with qualifizierter Hauptschulabschluss or less obtained at the Gymnasium. For most of these cases, the inconsistencies did not appear obvious enough to justify manipulation, so that only few edits were conducted.

A fourth step in the do-file deals with further inconsistencies related to obtainment of school-leaving degrees. First, when respondents were too young to plausible have attained a school-leaving degree, the value of the `schooldegree` variable was set to “no qualification intended”. Second, in 40 cases, lower-level school leaving degrees were recorded after the obtainment of higher-level school-leaving degrees. These cases have been corrected after case-wise inspection. Third, as a consequence of edits in the `schooldegree` variable, the variable `schoolgpa` was set to missing in cases where no school-leaving degrees have been obtained.

The do-file stores the edited dataset as `school_plausi.dta`. It also creates the following auxiliary variables:

`start`	start date of the spell (in months since January 1960)
`end`	end date of the spell (in months since January 1960)
`dur`	updated indicator of the total duration of the spell in months
`n`	consecutive enumeration of months within respondents
`startagem`	age at the start of the spell in months
`startagey`	age at the start of the spell in years
`endagem`	age at the end of the spell in months
`endagey`	age at the end of the spell in years

7.1.3 school_recode.do

The do-file `school_recode.do` starts from the cleaned and edited dataset `school_plausi.dta` and completes the recode of school trajectories and school-leaving degrees. The aim of this do-file is to create a variable containing detailed and monthly information on the highest school-leaving degree that the respondent possesses in a given month. It creates the following variable:

`schooldegree_cum` indicator of the highest school-leaving degree in a given month, coded:

0	“no degree”
1	“no degree or more (abroad/recog)”
2	“no degree or more (GER)”
3	“special needs (abroad/recog)”
4	“special needs (GER)”
5	“special needs or more (abroad/recog)”
6	“special needs or more (GER)”
7	“Haupt (abroad/recog)”
8	“Haupt (GER)”
9	“Haupt or more (abroad/recog)”
10	“Haupt or more (GER)”
11	“qual.Haupt (abroad/recog)”
12	“qual.Haupt (GER)”
13	“qual.Haupt or more (abroad/recog)”
14	“qual.Haupt or more (GER)”
15	“Mittl.Reife (abroad/recog)”
16	“Mittl.Reife (GER)”
17	“Mittl.Reife or more (abroad/recog)”
18	“Mittl.Reife or more (GER)”
19	“FH-Reife (abroad/recog)”
20	“FH-Reife (GER)”
21	“FH-Reife or more (abroad/recog)”
22	“FH-Reife or more (GER)”
23	“Abitur (abroad/recog)”
24	“Abitur (GER)”
.a	“indeterminate (abroad/recog)”
.b	“indeterminate (GER)”
.c	“rec.ind/abroad: Haupt”
.d	“rec.ind/abroad: qual.Haupt”
.e	“rec.ind/abroad: MR”
.f	“rec.ind/abroad: FHR”
.g	“rec.ind/abroad: Abitur”
.h	“rec.ind/abroad: indeterm.”

The variable follows a hierarchical structure, which is primarily defined by the hierarchy of the school-leaving degrees. Within each category, degrees obtained in Germany are assigned a higher hierarchy level than degrees that have been obtained abroad and have been recognized in Germany. Degrees that have been obtained abroad but have not been recognized in Germany are counted as “no degree”. If respondents obtain an additional degree with an indeterminate level, they are coded into the category “previous degree or more”. For example, if a respondent possesses a “Mittlere Reife” in month *t* and acquires an additional degree in month *t+1*, which is indeterminate, the respective value of the `schooldegree_cum` variable will be set to “Mittlere Reife or more” for month *t+1* and the following months until an additional higher-level degree shows up in a subsequent month. The missing codes .c to .h indicate cases with indeterminate records about their school-leaving degrees, but with information about school-leaving degrees that have been obtained abroad and have not been recognized in Germany. The do-file also stores some auxiliary variables that have been used in the construction of the `schooldegree_cum` variable:

schooldegree_add	indicator if an additional school-leaving degree has been obtained in the respective month
schooldegree2	recode of schooldegree, distinguishing between degrees obtained in Germany and abroad/recognized and coding degrees obtained abroad that have not been recognized in Germany as “no degree”
schooldta	indicator marking that the data is related to school-leaving degrees

The do-file stores the dataset as school_recode.dta. In addition, it also stores a compressed version of the dataset as school_recode_comp.dta, comprising only the lines of the full dataset when information on school-leaving degrees has changed.

7.2 Information on vocational degrees

7.2.1 voctrain_basics

The do-file voctrain_basics.do provides a first recode of relevant information related to vocational training or higher education. It starts from the month_episodes dataset and adds information from the original spVocTrain dataset, which contains detailed information about vocational training and higher education spells. In addition, it draws on information about the date of birth and the age of migration from the basicinfo dataset (see above) to help determine whether vocational training or higher education degrees have been obtained in Germany or abroad. It creates the following variables, which are stored in the dataset voctrain_basics.dta:

agem	age in months
agey	age in years
voctype	indicator of the type of vocational training or higher education attended
voctraincont	binary indicator whether the training episode still continues
vocdegree_detail	detailed indicator of the vocational degree obtained at the end of the episode, including categories coded as missings for: indeterminate, still attending this training and no qualification obtained
vocdegree	recoded version of vocdegree_detail, grouping its categories into: 1 VOC1 lower-level vocational training degree 2 VOC2 higher-level vocational training degree 3 HE1 higher education degree 4 HE2 post-higher education degree (e.g. doctorate, habilitation) 5 other qualification
vocgpa	indicator of grade point average (GPA) of the vocational degree on the German grading scale (1 very good – 6 insufficient). Note that this variable contains some values between 6 and 10, which stem from different grading scales, mostly used in particular higher education programs that have not been treated separately in the original questionnaires.
gertrain	binary indicator whether the vocational training or higher education program has been attended in Germany

fedtrain	indicator of the federal state, in which the vocational training or higher education program has been attended, including categories for: Germany (not further specified) and missing categories for: indeterminate, abroad

7.2.2 voctrain_plausi.do

The do-file voctrain_plausi.do conducts plausibility checks and data cleaning procedures to correct inconsistencies in the spell data on vocational training and higher education episodes. It starts from the voctrain_basics dataset. An issue in the spell data on vocational training episodes are parallel spells with duplicates. These redundant reports on the same training episodes most likely arise from data input queries in different waves as a result of problems in filtering.

Table 3

Example of a Case with Duplicates of Vocational Training Spells

ID_t	splink	voctype	vocdegree_detail	month	start	end	dur
8000530	240002	univ.appl. sciences (FH)	degree from a university of applied science	1971m4	1971m4	1974m1	34
8000530	240005	univ.appl. sciences (FH)	degree from a university of applied science	1971m4	1971m4	1974m3	36
8000530	240002	univ.appl. sciences (FH)	degree from a university of applied science	1971m5	1971m4	1974m1	34
8000530	240005	univ.appl. sciences (FH)	degree from a university of applied science	1971m5	1971m4	1974m3	36
...							

Table 3 provides an example of such a case and displays two months from a vocational training trajectory as recorded in the data. For each of the two months, we observe two parallel spells (240002 and 240005) at a university of applied sciences. Both spells complete with a degree from a university of applied sciences. Both spells have the same starting date, but end two months apart from each other (January and March 1974). While, in principle, it is possible that persons study two higher education programs at the same time, it appears more likely that this is a case of repeated recording of the same training episode in different data collection waves. The small difference in the end dates might be due to recall errors, which are typical for the collection of retrospective life course data with such a long distance between the event and the date of the survey. As a general rule for data edits, such parallel spells in identical training programs with the same vocational degree have been deleted if they took place in the same federal state and have been retained if they took place in different federal states. Since, for the initial purpose of this data cleaning process, a misclassification as duplicates or actual parallel training spells does not lead to any noteworthy bias, the cleaning of vocational training episodes was less detailed as the cleaning of school episodes. For other purposes, a more thorough data cleaning endeavor might be in order.

The do-file stores the edited dataset as voctrain_plausi.dta. It also creates the following auxiliary variables:

start	start date of the spell (in months since January 1960)
end	end date of the spell (in months since January 1960)
monthnr_voc	number of the spell within a given month, which can be larger than one if a respondent has multiple spells in the same month

7.2.3 voctrain_recode.do

The do-file voctrain_recode.do starts from the cleaned and edited dataset voctrain_plausi.dta and completes the recode of training trajectories and vocational degrees. The aim of this do-file is to create a variable containing detailed and monthly information on the highest vocational degree that the respondent possesses in a given month. It creates the following variable:

vocdegree_cumindicator of the highest school-leaving degree in a given month, coded:

0	"no vocational degree"
1	"no vocational degree or more (abroad)"
2	"no vocational degree or more (GER)"
3	"VOC1 – lower-level vocational degree (abroad)"
4	"VOC1 – lower-level vocational degree (GER)"
5	"VOC1 or more (abroad)"
6	"VOC1 or more (GER)"
7	"VOC2 – higher-level vocational degree (abroad)"
8	"VOC2 – higher-level vocational degree (GER)"
9	"VOC2 or more (abroad)"
10	"VOC2 or more (GER)"
11	"HE1 Higher education degree (abroad)"
12	"HE1 Higher education degree (GER)"
13	"HE1 or more (abroad)"
14	"HE1 or more (GER)"
15	"HE2 Post-higher education degree (abroad)"
16	"HE2 Post-higher education degree (GER)"
.a	"indeterminate (GER)"
.b	"indeterminate (abroad)"

Similar to the schooldegree_cum variable, the variable vocdegree_cum follows a hierarchical structure, which is primarily defined by the hierarchy of the vocational degrees. Within each category, degrees obtained in Germany are assigned a higher hierarchy level than degrees that have been obtained abroad. Unlike the corresponding variable indicating school-leaving degrees, this variable does not differentiate whether the vocational degrees have been recognized in Germany or not. The do-file also stores some auxiliary variables that have been used in the construction of the vocdegree_cum variable:

vocdegree_add	indicator if an additional vocational degree has been obtained in the respective month
vocdegree2	recode of vocdegree, distinguishing between degrees obtained in Germany or abroad
vocdta	indicator marking that the data is related to vocational degrees

The do-file stores the dataset as `voctrain_recode.dta`. In addition, it also stores a compressed version of the dataset as `voctrain_recode_comp.dta`, comprising only the lines of the full dataset when information on vocational degrees has changed.

7.3 Storing data on education in an information-by-age format

The final steps in the recode of information on educational trajectories and degrees consist in merging the data on school-leaving and vocational degrees and then transferring the data into an information-by-age format. The aim of this procedure is to provide data on education in the wide format, which means that the dataset contains one line for each respondent and age-specific variables with information on the highest school-leaving and vocational degrees at that given age.

7.3.1 educ_merge.do

The do-file `educ_merge.do` starts from the dataset `month_episodes.dta`, which contains information about respondents' states (school, vocational training, employment etc.) in each single month of their observation window. The aim of the do-file is to provide information about the highest school-leaving and vocational degrees that respondents possess in a given month. By merging information on respondents' date of birth from the dataset `basicinfo.dta`, respondents' age can be determined for each single month (c.f *Table 4* with information on age in months and age in years in the last two columns).

Table 4

Adding Age Information to the Information-by-Month Dataset

ID_t	month	birthdate	agem	agey
8000215	1966m7	1960m7	72	6
8000215	1966m8	1960m7	73	6.083333
8000215	1966m9	1960m7	74	6.166667
8000215	1966m10	1960m7	75	6.25
8000215	1966m11	1960m7	76	6.333333
8000215	1966m12	1960m7	77	6.416667
8000215	1967m1	1960m7	78	6.5
8000215	1967m2	1960m7	79	6.583333
8000215	1967m3	1960m7	80	6.666667
8000215	1967m4	1960m7	81	6.75

While the `month_episodes.dta` data can contain respondents with multiple states per month (cf. *Table 1*), this information will not be relevant for the further recode. Hence, all lines with parallel states are deleted, so that for each respondent only one line per month/age remains in the dataset.

The calculation of the age for each month of a respondent's observation window allows for a plausibility inspection of the resulting age values. As the data query of educational trajectories in the NEPS Starting Cohort 6 typically starts with elementary school episodes, one would expect the lowest age values around the age of six. *Table 5* shows the actual ages that can be found in the data by categories. For about 88 percent of the respondents, the spell data only

comprises months when they were 6 years or older. For about 8.5 percent of the respondents, the data comprises spells when they were between 5.5 and 6 years old and for about 2 percent of the respondents the youngest age recorded is between 5 and 5.5 years. While these age groups might still be plausible if the respondents started elementary school unusually early, the remaining categories reflect highly implausible age values. For about 1.2 percent of the respondents, the data contains spells when they were younger than 5 years old. For nine respondents, the data even contains negative age values. These implausible values most likely reflect input errors regarding either the year of birth or the timing of the school episodes. At this stage, these respondents with implausible age values remain in the data. They can, however, be identified with the variable `lowage`.

Table 5

Plausibility Inspection of Age Values (Tabulation of `lowage`)

Respondent's observation window includes	**Frequency**	**Percent**	**Cumulative**
0. only months with age >= 6	15,118	88.21	88.21
1. months with 5.5 <= age < 6	1,459	8.51	96.72
2. months with 5 <= age < 5.5	340	1.98	98.70
3. months with 0 <= age < 5	213	1.24	99.95
4. months with age < 0	9	0.05	100.00
Total	17,139	100.00	

In a next step, information on school-leaving degrees are added to the dataset by merging information from `school_recode.dta`. Since this dataset only contains information on school episodes, the resulting data contains gaps in the variable indicating the highest school-leaving degree that a respondent possesses in a given month (`schooldegree_cum`) for episodes other than school episodes. The upper panel of *Table 6* shows such an example. The variable `schooltype` indicates that the respondent attended a lower secondary school until August 1981 and obtained a "Mittlere Reife" in this month. After an episode in vocational training, the respondent attended the Gymnasium until June 1985 and obtained a "Fachhochschulreife" there. Hence, the variable `schooldegree_cum` should indicate "Mittlere Reife" for all months between August 1981 and Mai 1985 and "Fachhochschulreife" for all months after June 1985. Since the months following the school episodes are in the states "vocational training" and "employment", they were not included in the dataset `school_recode.dta`, which is why the merging procedure generated missing values in the `schooldegree_cum` variable. The do-file adjusts for this by filling up all subsequent lines with the same school-leaving degree until the respondent has a record of a higher-level school-leaving degree (cf. lower panel of *Table 6*). For respondents who never obtained any school-leaving degree, the do-file assigns the missing code .m "case without any schooldegree information" in the `schooldegree_cum` variable.

The next step adds information on vocational degrees to the dataset by merging information from `voctrain_recode.dta`. Like in the example before, empty lines in the `vocdegree_cum` variable are filled up, so that for each month, the variable indicates the highest vocational degree that the respondent possesses in that moment. For respondents

who never obtained any vocational degree, the do-file assigns the missing code .m “case without any vocdegree information” in the `vocdegree_cum` variable.

The do-file also expands the time-constant information on the age of migration from `voctrain_recode.dta` to all months and saves the data as `educ_merge.dta`

Table 6

Gaps in schooldegree_cum Variable, Example Before and After Fill-up Procedure

Before fill-up procedure				
ID_t	month	sptype	schooltype	schooldegree_cum
8000338	1981m6	School	Haupt/Real/POS	no degree
8000338	1981m7	School	Haupt/Real/POS	no degree
8000338	1981m8	School	Haupt/Real/POS	Mittl.Reife (GER)
8000338	1981m9	VocTrain	.	.
8000338	1981m10	VocTrain	.	.
8000338	1981m11	VocTrain	.	.
...				
8000338	1985m4	School	Gymnasium	Mittl.Reife (GER)
8000338	1985m5	School	Gymnasium	Mittl.Reife (GER)
8000338	1985m6	School	Gymnasium	FH-Reife (GER)
8000338	1985m7	Emp	.	.
8000338	1985m8	Emp	.	.
8000338	1985m9	Emp	.	.
After fill-up procedure				
ID_t	month	sptype	schooltype	schooldegree_cum
8000338	1981m6	School	Haupt/Real/POS	no degree
8000338	1981m7	School	Haupt/Real/POS	no degree
8000338	1981m8	School	Haupt/Real/POS	Mittl.Reife (GER)
8000338	1981m9	VocTrain	.	Mittl.Reife (GER)
8000338	1981m10	VocTrain	.	Mittl.Reife (GER)
8000338	1981m11	VocTrain	.	Mittl.Reife (GER)
...				
8000338	1985m4	School	Gymnasium	Mittl.Reife (GER)
8000338	1985m5	School	Gymnasium	Mittl.Reife (GER)
8000338	1985m6	School	Gymnasium	FH-Reife (GER)
8000338	1985m7	Emp	.	FH-Reife (GER)
8000338	1985m8	Emp	.	FH-Reife (GER)
8000338	1985m9	Emp	.	FH-Reife (GER)

7.3.2 educ_age.do

The do-file `educ-recode.do` conducts some final recodes based on `educ_merge.dta` and transfers the data into an information-by-age format. Since the focus of this recode is on the age at which respondents have obtained their school-leaving and vocational degrees, another age-related plausibility inspection is in order. *Table 7* shows an excerpt of a tabulation

of the age at which respondents obtained their first school-leaving (left-hand side) and vocational (right-hand side) degrees. The table documents that with regard to both school-leaving and vocational degrees there are some respondents in the dataset with implausible information on the age at which they obtained these degrees. For the birth cohorts in the dataset, the lowest age at which the first school-leaving degrees can be obtained should be around the age of 14-15 (persons with "Hauptschulabschluss"). Yet, about 3 percent of the respondents have data entries suggesting they obtained their first school-leaving degrees between age 13 and 14. While this might still be acceptable as outliers with an unusually low age, about 0.9 percent of the respondents have a record of an age of 13 or younger in the dataset. These, and, in particular, the three respondents with negative age values point to data entry errors related to either the data of birth or the timing of the school trajectories. With regard to vocational degrees, the share of outliers with implausible age values at their first vocational degree is somewhat lower. While the lowest age for vocational degrees might be expected around the age of 15-16, only about 0.1 percent of all respondents recorded an age of 14 or younger. At this stage, these respondents with implausible age values remain in the data. They can, however, be identified with the variables `schooldegree_1_age` and `vocdegree_1_age`, which provide the age of the first school-leaving and vocational degrees as time-constant information for each respondent.

Table 7

Age at First School-leaving and First Vocational Degree

age first school-leaving degree				**age first vocational degree**			
	Freq.	Percent	Cum.		Freq.	Percent	Cum.
-12.33	1	0.01	0.01	-4.67	1	0.01	0.01
-8.58	1	0.01	0.01	-2.00	1	0.01	0.01
-0.5	1	0.01	0.02	4.17	1	0.01	0.02
5.67	1	0.01	0.02	6.00	1	0.01	0.02
7	1	0.01	0.03	8.5	1	0.01	0.03
7.25	1	0.01	0.04	8.67	1	0.01	0.04
8.00	1	0.01	0.04	9.67	1	0.01	0.04
10-11	5	0.05	0.09	10-11	3	0.03	0.07
11-12	26	0.16	0.25	11-12	2	0.01	0.08
12-13	115	0.68	0.93	12-13	0	0.00	0.08
13-14	530	3.09	4.02	13-14	5	0.05	0.13
...				14-15	20	0.14	0.27
				15-16	90	0.51	0.78
				...			

The central step in the do-file is the transfer of monthly data into an information-by-age format. Based on version 14-0-0 of the NEPS Starting Cohort 6, the do-file generates information on school-leaving and vocational degrees from age 10 to age 77. The dataset covers 17,139 respondents (including some respondents with implausible age values). Since an age-year covers 12 months during which information on educational degrees can change (e.g. a person can acquire a new school-leaving degree four months after his or her 16th birthday), it is vital to define from which of these months the respective degrees will be recorded. This do-file records information from the first month of a given age-year. For

example, if a person turns 30 in October of a given year, the variables indicating school-leaving and vocational degrees at age 30 refer to the highest school-leaving and vocational degrees that this person possessed in October of that year.

Table 8

Distribution of School-leaving Degrees at Different Ages (N=17,139)

highest school leaving degree at age	10	20	40	60	77
0 no degree	99.72	4.86	2.98	1.17	
1 no degree or more (abroad)	0.01	1.63	1.28	0.40	
2 no degree or more (GER)	0.02	0.26	0.19	0.06	
3 special needs (abroad+recog)					
4 special needs (GER)		0.30	0.19	0.03	
5 special needs or more (abroad)					
6 special needs or more (GER)			0.01		
7 Haupt (abroad+recog)		0.61	0.46	0.20	
8 Haupt (GER)	0.02	17.24	14.80	8.47	
9 Haupt or more (abroad)		0.03	0.03		
10 Haupt or more (GER)		0.34	0.34	0.25	
11 qual.Haupt (abroad+recog)		0.15	0.13	0.06	
12 qual.Haupt (GER)		5.41	4.10	0.79	
13 qual.Haupt or more (abroad)		0.01	0.01		
14 qual.Haupt or more (GER)		0.09	0.08	0.02	
15 Mittl.Reife (abroad+recog)		0.96	0.72	0.23	
16 Mittl.Reife (GER)	0.01	35.71	28.04	10.05	0.02
17 Mittl.Reife or more (abroad)		0.05	0.05	0.01	
18 Mittl.Reife or more (GER)		0.57	0.59	0.20	
19 FH-Reife (abroad+recog)		0.08	0.09	0.02	
20 FH-Reife (GER)		4.11	4.94	1.46	
21 FH-Reife or more (abroad)			0.01	0.01	
22 FH-Reife or more (GER)			0.05	0.02	
23 Abitur (abroad+recog)		1.11	0.92	0.39	0.01
24 Abitur (GER)	0.01	26.29	22.29	7.27	
.m case without any schooldegree info	0.20	0.20	0.15	0.06	
.o out of empirical data range	0.02		17.57	68.83	99.98

As it is implausible that respondents have obtained any school-leaving or vocational degree before the age of 10, the do-file does not generate variables indicating degrees for those age groups. *Table 8* displays the distributions of five selected variables `schooldegree_10`, `schooldegree_20`, `schooldegree_40`, `schooldegree_60`, `schooldegree_77`, referring to school-leaving degrees at ages 10, 20, 40, 60, and 77. At the age of 10, except for some outliers with implausible ages for their first school-leaving degrees (as indicated by variable `schooldegree_1_age`), almost all respondents do not possess any school-leaving degree. The variable contains the missing codes `.m` for respondents without any information on school-leaving degrees and `.o` for respondents, for whom the age is outside the observable data range. In the case of `schooldegree_10`, 34 respondents (0.02 percent) display the

missing code .o, which is due to the fact that the recorded age range of their spells ends before the age of ten. This is either due to non-completed interviews or due to data errors in the time-lines of their life-course data (cf. variable lowage). At the age of 20, almost all respondents have obtained at least one school-leaving degree. Only 4.9 percent of the respondents do not possess any degree. Throughout the variables indicating school-leaving degrees at the ages of 40 and 60, an increasing fraction of the respondents is outside the observable data range. This can have two reasons. The first reason is panel attrition, when respondents have reached the respective age during the survey, but did not participate. The second reason is the age of the respondents. For example, the highest school-leaving degree at the age of 60 can only be observed for persons who are 60 or older at the time of the interview. Since the data covers birth cohorts from 1944 to 1986, the younger cohorts are not represented in variables indicating school-leaving degrees at higher ages. The highest age that can be observed in the data is 77 years (represented by four respondents).

The do-file stores the following variables in the dataset educ_age.dta:

lowage	time-invariant categorical variable indicating respondents with unusually low age in their early spells (cf. *Table 5*). This variable can be used to exclude respondent with implausible age values
schooldegree_1_age	time-invariant variable indicating respondents' age at obtainment of the first school-leaving degree. This variable can be used to exclude respondents with implausibly low or high age
vocdegree_1_age	time-invariant variable indicating respondents' age at obtainment of the first school-leaving degree. This variable can be used to exclude respondents with implausibly low or high age
schooldegree_*	indicator of respondents' highest school-leaving degree at ages 10 to 77.
schooldegree_abroad_*	indicator whether the highest school-leaving degree recorded in variables schooldegree_10 - schooldegree_77 has been acquired in Germany or abroad
schooldegree_add_*	indicator whether at the given age (from 10 to 77) a new higher-level school-leaving degree has been acquired
schoolgpa_*	GPA of the highest school-leaving degree recorded in variables schooldegree_10 - schooldegree_77
vocdegree_*	indicator of respondents' highest vocational degree at ages 10 to 77.
vocdegree_abroad_*	indicator whether the highest vocational degree recorded in variables vocdegree_10 - vocdegree_77 has been acquired in Germany or abroad
vocdegree_add_*	indicator whether at the given age (from 10 to 77) a new higher-level vocational degree has been acquired

8. Recode of labor market information

The recode of labor market information comprises three sets of do-files. The first one prepares information on earnings trajectories. It consists of a thorough plausibility inspection and correction of information on earnings from gainful employment and a transfer of the data into the information-by-age format. The second one comprises a recode of information on characteristics of respondents' occupations and, again, the transfer of the data into an information-by-age format. The third one generates information on the date of respondents' first significant employment.

8.1 Information on earnings

8.1.1 income_plausi.do

The do-file `income_plausi.do` conducts a plausibility inspection and data cleaning of the earnings data. It starts from the original `spEmp` data, which provides information on employment for each survey wave. It can contain multiple employment spells per wave if respondents had more than one employment. In each wave, respondents are asked to state their gross and net earnings from the previous month and their annual bonus payments of the previous year. Hence, information on earnings are only available in the panel component of the NEPS Starting Cohort 6, but not for retrospective employment biographies. Since earnings data from the first wave (ALWA sample) is not entirely comparable to information from subsequent waves (NEPS waves), the do-file deletes earnings information from the first wave. To assign the exact month to which earnings information refer to, the do-file merges information on the dates of the interviews from the `CohortProfile` data. It also merges information on the start and end dates of the employment spells from the `Biography` data. This only serves facilitation of the visual data inspection and the do-file does not draw on this information for the recode process.

Table 9

Mean and Median Gross Monthly Earnings, by Earnings Categories

ts23511_g1	Mean	Median
less than 500 euros	234.04	230
500 to less than 1,000 euros	718.19	715
1,000 to less than 1,500 euros	1,227.33	1,200
1,500 to less than 2,000 euros	1,703.30	1,700
2,000 to less than 2,500 euros	2,207.44	2,200
2,500 to less than 3,000 euros	2,683.68	2,700
3,000 to less than 4,000 euro	3,412.10	3,400
4,000 to less than 5,000 euros	4,379.62	4,400
5,000 euros and more	7,676.75	6,200
Total	2,954.16	2,500

A first substantive recode in the do-file concerns the handling of missing data in the variables based on an open query of earnings. Respondents who do not provide earnings information in the open query are filtered into a series of queries with predefined earnings categories. If respondents provide information based on the categorical query but not in the open question,

the do-file replaces their earnings values in the open variable with the respective mean values of respondents who did provide open figures about their earnings. *Table 9* provides an example for gross monthly earnings. As the mean and median values are not far apart, the do-file used mean values for the replacement of missing information in the open query. The same procedure is applied to net monthly earnings.

The next step in the do-file is concerned with cleaning of data duplicates. *Table 10* shows an example of a case with multiple spells per wave (as indicated by the `n` variable) and a duplicate (as indicated by the `dup1g` variable). To distinguish between actual multiple spells and duplicates, the do-file applies a series of duplicate checks based on the variables splink, wave, start, end, and variables indicating working hours and earnings.

Table 10

Example of Multiple Spells and Duplicates in the Employment Data

ID_t	splink	wave	start	end	ts23510	ts23223_g1	n	dup1g
...								
8000335	260005	5	2005m9	2017m4	2400	20	1	0
8000335	260006	5	2010m11	2020m10	800	10	2	0
8000335	260005	6	2005m9	2017m4	.b	10	1	0
8000335	260006	6	2010m11	2020m10	800	.x	2	0
8000335	260005	8	2005m9	2017m4	1250	.x	1	0
8000335	260006	8	2010m11	2020m10	1000	.x	2	0
8000335	260006	9	2010m11	2020m10	9000	25	1	0
8000335	260006	10	2010m11	2020m10	10000	20	1	0
8000335	260006	11	2010m11	2020m10	12000	18	1	0
8000335	260006	12	2010m11	2020m10	15000	11	1	1
8000335	260006	12	2010m11	2020m10	15000	11	2	1

In the example displayed in *Table 10*, the inspection shows that the multiple spells in waves 5, 6 and 8 clearly stem from different parallel employment activities as the start and end dates, the spell indicator `splink` and the earnings differ. Instead, the multiple spells in wave 12 most likely reflect a data duplicate as the values in all variables are identical.

The final step in the do-file inspects and corrects unusually high or low values for earnings or working hours. To this end, it generates auxiliary variables indicating the cumulative net and gross monthly earnings (`temp4_total`, column 7 in *Table 11*) and the cumulative working hours (`temp4_wh`, last column in *Table 11*) per wave. As a general rule, the do-file recodes all cases as missing that have earnings or working hours values that deviate by the factor 2 from the respective earnings or working hours values in the same spell from previous or later waves. For the remaining outliers, the do-file conducts a cleaning procedure of implausible earnings and working hour information based on a thorough case-wise inspection.

Table 11

Example of Unusually High Values for Earnings and Working Hours

ID_t	wave	splink	start	end	gross_~n	temp4_~l	ts2~3_g1	temp4_wh
...								
8000671	4	260003	2009m7	2012m2	4750	4750	55	55
8000671	5	260006	2012m3	2018m12	4500	4500	60	60
8000671	6	260006	2012m3	2018m12	4500	6000	70	77
8000671	6	260007	2013m3	2015m2	1500	6000	7	77
8000671	7	260006	2012m3	2018m12	8000	15676	60	120
8000671	7	260007	2013m3	2015m2	7676	15676	60	120
8000671	8	260006	2012m3	2018m12	5700	5700	70	70
8000671	9	260006	2012m3	2018m12	5500	5500	65	65
8000671	10	260006	2012m3	2018m12	5200	5200	70	70

The example displayed in *Table 11* shows a rather complicated case with unusually high earnings and working hours in wave 7 as the result of two overlapping employment spells with high earnings and working hours each. A closer inspection with the consultation of further variables did not hint at any obvious error in the data, which is why the case has not been edited further and remained in the dataset as is. Such case-wise inspections have been undertaken for more than 100 individual cases with conspicuous values on earnings or working hours.

The do-file saves the dataset with cleaned earnings and working hours as `income_plausi.dta`.

8.1.2 income_age.do

The do-file `income_age.do` conducts further recodes of the data stored in `income_plausi.dta` and transfers it into an information-by-age format.

One substantive data manipulation concerns the handling of missing information related to gross annual bonus payments. In each wave, respondents were asked whether they received bonus payments in the previous year in six different categories (such as 13th month salary, Christmas bonus, holiday bonus etc.). For each category, for which they stated to have received a bonus payment, they were asked about the exact gross amount of the payment. In some cases, respondents stated that they received a bonus payment but did not provide the amount. The do-file conducts a simple imputation procedure for those cases. For each of the six bonus payment categories, it calculates the share of the bonus payment relative to the gross monthly earnings for each respondent with a respective bonus payment. Then, it calculates the average of the individual relative shares of bonus payments within each earnings category of the variable `ts23511_g1`.

Table 12 shows the average relative shares of holiday payments by categories of gross monthly earnings as an example. The average holiday payments vary between 29 and 59 percent over the categories of gross monthly earnings. Within these categories of variable `ts23511_g1`, the do-file replaces missing values on bonus payments with the respective average relative value of gross monthly earnings, separately for all six types of bonus payments.

Table 12

Average Share of Holiday Bonuses relative to Gross Monthly Earnings by Earnings Categories

ts23511_g1	Mean	Freq.
less than 500 euros	.59303546	15,974
500 to less than 1,000 euros	.38883328	6,071
1,000 to less than 1,500 euros	.33814102	7,384
1,500 to less than 2,000 euros	.29415748	8,458
2,000 to less than 2,500 euros	.31375995	10,037
2,500 to less than 3,000 euros	.31199238	9,376
3,000 to less than 4.000 euro	.31226149	16,111
4,000 to less than 5,000 euros	.34228086	10,229
5,000 euros and more	.34969044	15,341
Total	.37168405	98,981

The next step in the do-file concerns the cumulation of earnings and working hours per wave. *Table 13* provides an example for gross monthly earnings and actual weekly working hours. In waves 4 to 7, the respondent reported two different employment spells. The gross monthly earnings and the weekly working hours for each single spell are recorded in the variables `gross_earn` and `workhours`. The cumulated earnings and working hours for each wave are stored in the variables `gross_earn_c` and `workhours_c`. In those cases, when one of the single spells contains missing information, the variables indicating the cumulative earnings or working hour values are set to missing as well. At this stage, the data is still in long format. The do-file saves this intermediate step as `income_age_long.dta`.

Table 13

Cumulated Earnings and Working Hours by Wave

ID_t	wave	n	splink	gross_earn	gross_earn_c	workhours	workhours_c
8000335	2	1	260005	400	400	6	6
8000335	3	1	260005	1500	1500	22	22
8000335	4	1	260006	750	2550	10	25
8000335	4	2	260005	1800	2550	15	25
8000335	5	1	260005	2207	.a	25	37
8000335	5	2	260006	.b	.a	12	37
8000335	6	1	260005	2400	3200	20	30
8000335	6	2	260006	800	3200	10	30
8000335	7	1	260005	718	1518	10	.a
8000335	7	2	260006	800	1518	.a	.a

The final step in the do-file consists of the transfer of the data into the information-by-age format. For that, it merges information on the date of birth from `basicinfo.dta` and generates a variable indicating the age in years when earnings have been reported. The do-file rounds the variable to its floor, reflecting completed full years. In some cases, when the distance between two interviews in consecutive survey waves was less than a year and close to the respondent's birthday, this procedure creates two data entries for the same age and a

gap for either the previous or the subsequent age. In the latter cases, the older age from the double entry is rounded to the next integer to fill the age gap. In all other cases with double age entries, information with the highest earnings values is retained and the information with the lowest values deleted.

The do-file stores the following variables in the dataset `earnings_age.dta`:

`earnage_min`	time-invariant variable indicating respondents' youngest age with information about earnings
`earnage_max`	time-invariant variable indicating respondents' oldest age with information about earnings
`jobs_*`	indicator of respondents' number of parallel jobs at ages 21 to 76.
`gross_earn_*`	indicator of respondents' gross monthly earnings from gainful employment at ages 21 to 76
`net_earn_*`	indicator of respondents' net monthly earnings from gainful employment at ages 21 to 76
`gross_bonus_*`	indicator of respondents' annual bonus payments from gainful employment at ages 21 to 76
`gross_bonus_miss_*`	binary indicator whether respondents' information on annual bonus payments at ages 21 to 76 contains missing information
`workhours_*`	indicator of respondents' actual total weekly working hours at ages 21 to 76

8.2 Information on respondents' occupation

The do-file `occupation_age.do` recodes information on respondents' employment history and characteristics of the occupations held and transfers the data into an information-by-age format. It is based on the dataset `month_episodes.dta` and merges information on the date of birth from `basicinfo.dta`. Since the information on employment histories also stems from the retrospective survey of the NEPS, Staring Cohort 6, it covers an age range from 10 to 77, like the data on education. Unlike the information on earnings or education, the information on employment histories are not further plausibility checked. The do-file stores the following variables in the dataset `occupation_age.dta`:

`empage_min`	time-invariant variable indicating respondents' youngest age with information about employment. Note that this variable contains some cases with implausibly low values for the same reasons as stated above in section 7.3.2
`empage_max`	time-invariant variable indicating respondents' oldest age with information about employment.
`status_*`	indicator of respondents' employment status at ages 10 to 77. Based on information from the parallel spells, the variable also contains categories

	indicating parallel status, such as “employed and education”, including implausible combinations, such as “employed and unemployed”.
`egp_*`	indicator of respondents’ EGP class (11-category version) at ages 10 to 77. In case of multiple jobs, the variable displays the highest value
`isei_*`	indicator of respondents’ International Socio-Economic Index (ISEI) value at ages 10 to 77 (based on ISCO-08). In case of multiple jobs, the variable displays the highest value
`profstatus_*`	indicator of respondents’ professional status at ages 10 to 77
`pubserv_*`	indicator whether the respondent is employed in the private or public sector at ages 10 to 77
`fedwork_*`	indicator of the federal state of the work place at ages 10 to 77
`emppara_*`	indicator of respondents’ average number of parallel monthly employment spells at ages 10 to 77
`empmonths_*`	indicator of respondents’ number of months in employment at ages 10 to 77
`workh_c_*`	indicator of respondents’ average contractual weekly working hours at ages 10 to 77. Note that, unlike the variable `workhours` stored in the dataset `earnings_age.dta`, this variable has not been plausibility checked and edited
`workh_a_*`	indicator of respondents’ average actual weekly working hours at ages 10 to 77. Note that, unlike the variable `workhours` stored in the dataset `earnings_age.dta`, this variable has not been plausibility checked and edited

8.3 Information on respondents’ first significant employment

The do-file `date_employ1.do` starts from the dataset `month_episodes.dta`. It merges information from the original `spEmp` data and information on respondents’ date of birth from `basicinfo.dta` to recode information on the date and age of respondents’ first significant employment. The date of the first significant employment is supposed to indicate when respondents have entered the labor market with employment being their main activity. Respondents might have had previous employment episodes next to other main activities, such as students jobs next to studying or short-term employment to bridge two episodes in other activities (e.g. jobs during school vacations). Such minor episodes in employment are disregarded when constructing the date of entry into the first significant employment. The following criteria are applied to define significant employment: First, the cumulative working time in the jobs of the employment episode has to amount to at least 20 hours per week. Second, the employment episode has to include a job with an uninterrupted duration of at least 12 months without any parallel states in education, vocational training, military service or parental leave.

The do-file recodes the age (both in months and in full years) at the start of the first significant employment. In some cases, this leads to implausible age values (cf. *Table 14*). This is either a

consequence of biased information on the date of birth or of biased information on the timing of the employment spells. As the table shows, this only concerns a very small number of respondents. Note, however, that these cases remain in the data and have to be cleaned in a later stage if requested.

Table 14

Implausible Values in Variable Indicating Age at the First Significant Employment

age first significant job (in years)	Freq.	Percent	Cum.
-2	1	0.01	0.01
4	1	0.01	0.01
5	1	0.01	0.02
8	1	0.01	0.02
9	3	0.02	0.04
10	3	0.02	0.06
11	4	0.02	0.08
12	8	0.05	0.13
13	24	0.14	0.27
14	106	0.62	0.89
15	159	0.93	1.81
...			

The do-file stores the data in wide format in the dataset `date_employ1.dta`. It comprises the following variables:

`everemp`	binary variable indicating whether the respondent has any employment spells in the data
`eversigjob`	binary variable indicating whether the respondent has any record of a significant employment during the observation window
`datesigjob1`	date of the beginning of the first significant employment (in months since January 1960)
`yearsigjob1`	year of the beginning of the first significant employment
`agesigjobm`	age at the beginning of the first significant employment (in months)
`agesigjoby`	age at the beginning of the first significant employment (in full years)

9. Merge data

The do-file `merge_age.do` conducts the final step in the recoding procedure and merges information from the different previously constructed datasets into one single file. It compiles life course data in the information-by-age format from `educ_age.dta`, `occupation_age.dta` and `income_age.dta`. It also adds time-constant information from `basicinfo.dta`, `wavepart.dta` and `date_employ1`.dta.

Storing the data in an information-by-age format implies that respondents have missing values on age-specific variables when their age at the last wave they participated is lower than the age specified in the age-specific variable. For example, persons born in 1980 who participated in all waves until 2022 (wave 14) have a maximum observable age of 42. For variables indicating education or employment information at later ages, they have missing values by design. To indicate these missing values that lie outside the theoretical age-related observation window, all age-specific variables are assigned the missing code `.p "out of age range"`. The observable age range is even narrower for persons who dropped out of the survey before the last wave. For respondents who theoretically lie within the observable age range, but dropped out of the sample, all age-specific variables are assigned the missing code `.o "out of observation window"`.

Based on version 14 of the Starting Cohort 6, the recoded dataset comprises age-specific variables covering an age range from 10 to 77. An exception are the variables from the `income_age` dataset, which cover an age range from 21 to 76. The reason is that this information comes from the prospective panel, which has an age minimum of 21. As information on earnings refer to the year prior to the survey wave, the data does not include earnings information for age 77 even though some respondents reached that age. As earnings information come from the prospective panel component, this has also implications for missing information due to lying outside the theoretical age range. While all retrospective information is right-truncated only, information from the prospective panel can be both left- and right-truncated. *Figure 1* exemplifies this.

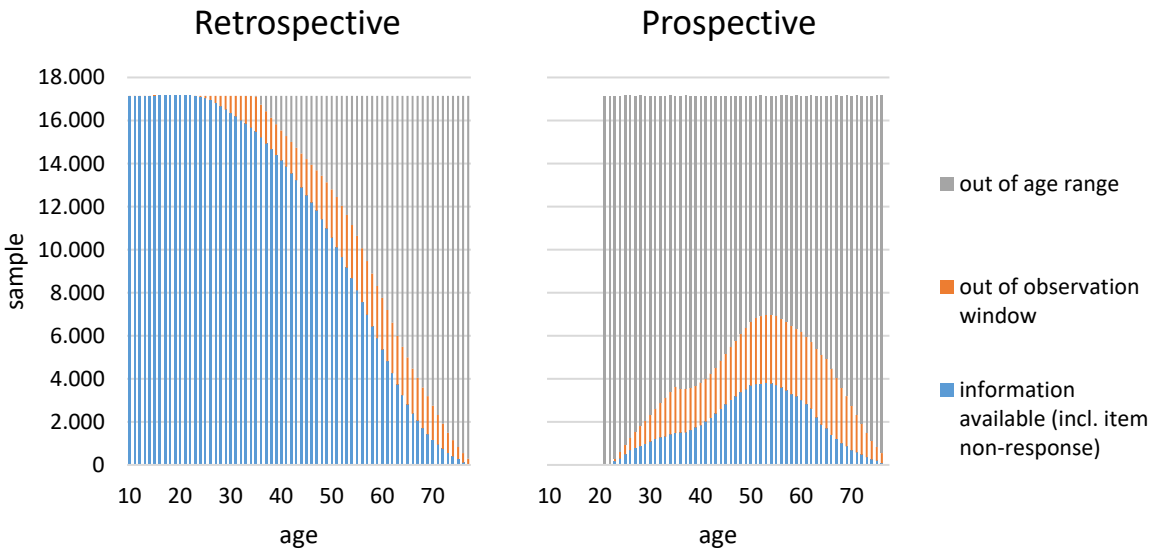

Figure 1. Missing information due to unobserved ages and panel dropout.

The figure shows the systematic missingness due to unobserved age-specific information, which arises from early panel dropout (orange) or age truncation (grey). The left-hand side refers to all data that can rely on retrospective information, such as education or employment histories. Here, the missing data leads to right-censoring only. The right-hand side of the figure refers to earnings information from the prospective panel component. As respondents enter the survey at different ages, the information-by-age format results both in left- and right-censoring of the age-specific data. As individuals have a maximum observation window of 12

years (from wave 2 to wave 14), information on age-specific earnings only overlaps for a small fraction of the sample. The distribution of available earnings information (including missings due to item-nonresponse) peaks at the age of 53 with information from persons born between 1955 and 1969. Excluding missings due to item-nonresponse, information on gross earnings are available for 3,395 respondents and information on net earnings for 3,417 respondents at age 53.

The do-file stores the dataset as `infobyage.dta`. *Table 15* provides the complete list of variables included in the dataset.

Table 15

List of variables included in the dataset `infobyage.dta`

Time-invariant information	
`ID_t`	personal identifier of respondent
`part_w1 - part_w14`	binary indicator of respondents participated in waves 1-14
`date_w1 - date_w14`	date of the interview in waves 1-14 (in months since January 1960)
`lastint`	date of respondents' last interview (in months since January 1960)
`partfirst`	indicator of the first wave of respondents' participation
`partlast`	indicator of the last wave of respondents' participation
`partnum`	number of waves that respondents have participated
`age_w2m`	respondents' age in wave 2 in months
`age_w2y`	respondents' age in wave 2 in full years
`age_w14m`	respondents' age in wave 14 in months
`age_w14y`	respondents' age in wave 14 in months
`birthdate`	respondents' date of birth (in months since January 1960)
`birthyear`	respondents' year of birth
`gender`	indicator of the respondent's gender
`casmin_f`	indicator of father's education based on the CASMIN classification
`casmin_m`	indicator of mother's education based on the CASMIN classification
`casmin_p`	indicator of parents' highest education based on the CASMIN classification
`educyears_f`	indicator of father's years of education derived from the typical duration of his education degrees

Table 15, continued

educyears_m	indicator of mother's years of education derived from the typical duration of her education degrees
educyears_p	indicator of parents highest years of education derived from the typical duration of their education degrees
egp_f	indicator of father's social class based on the EGP class schema (11-category version)
egp_m	indicator of mother's social class based on the EGP class schema (11-category version)
egp_p	indicator of parents' highest social class based on the EGP class schema (11-category version)
isei_f	indicator of father's social status based on the International Socio-Economic Index of Occupational Status (ISEI, ISCO-08 version)
isei_m	indicator of mother's social status based on the International Socio-Economic Index of Occupational Status (ISEI, ISCO-08 version)
isei_p	indicator of parents' highest social status based on the International Socio-Economic Index of Occupational Status (ISEI, ISCO-08 version)
siops_f	indicator of father's social prestige based on the Standard International Occupational Prestige Scale (SIOPS, ISCO-08 version)
siops_f	indicator of mother's social prestige based on the Standard International Occupational Prestige Scale (SIOPS, ISCO-08 version)
siops_p	indicator of parents' highest social prestige based on the Standard International Occupational Prestige Scale (SIOPS, ISCO-08 version)
migration	binary indicator whether respondents have a migration history
migration_gen	indicator of the migration generation based the NEPS migration generation concept
migration_age	indicator of the age of immigration (in years) for first and first-and-a-half generation immigrants
lowage	time-invariant categorical variable indicating respondents with unusually low age in their early spells (cf. Table 5). This variable can be used to exclude respondent with implausible age values

Table 15, continued

schooldegree_1_age	time-invariant variable indicating respondents' age at obtainment of the first school-leaving degree. This variable can be used to exclude respondents with implausibly low or high age
vocdegree_1_age	time-invariant variable indicating respondents' age at obtainment of the first school-leaving degree. This variable can be used to exclude respondents with implausibly low or high age
everemp	binary variable indicating whether the respondent has any employment spells in the data
eversigjob	binary variable indicating whether the respondent has any record of a significant employment during the observation window
datesigjob1	date of the beginning of the first significant employment (in months since January 1960)
yearsigjob1	year of the beginning of the first significant employment
agesigjob1m	age at the beginning of the first significant employment (in months)
agesigjob1y	age at the beginning of the first significant employment (in full years)
empage_min	time-invariant variable indicating respondents' youngest age with information about employment. Note that this variable contains some cases with implausibly low values
empage_max	time-invariant variable indicating respondents' oldest age with information about employment.
earnage_min	time-invariant variable indicating respondents' youngest age with information about earnings
earnage_max	time-invariant variable indicating respondents' oldest age with information about earnings
Age-specific information including retrospective data	
schooldegree_10 - schooldegree_77	indicator of respondents' highest school-leaving degree at ages 10 to 77
schooldegree_abroad_10 - schooldegree_abroad_77	indicator whether the highest school-leaving degree recorded in variables schooldegree_10 - schooldegree_77 has been acquired in Germany or abroad
schooldegree_add_10- schooldegree_add_77	indicator whether at the given age (from 10 to 77) a new higher-level school-leaving degree has been acquired

Table 15, continued

schoolgpa_10 - schoolgpa_77	GPA of the highest school-leaving degree recorded in variables schooldegree_10 - schooldegree_77
vocdegree_10 - vocdegree_77	indicator of respondents' highest vocational degree at ages 10 to 77
vocdegree_abroad_10 - vocdegree_abroad_77	indicator whether the highest vocational degree recorded in variables vocdegree_10 - vocdegree_77 has been acquired in Germany or abroad
vocdegree_add_10 - vocdegree_add_77	indicator whether at the given age (from 10 to 77) a new higher-level vocational degree has been acquired
status_10 - status_77	indicator of respondents' employment status at ages 10 to 77. Based on information from the parallel spells, the variable also contains categories indicating parallel status, such as "employed and education", including implausible combinations, such as "employed and unemployed"
egp_10 - egp_77	indicator of respondents' EGP class (11-category version) at ages 10 to 77. In case of multiple jobs, the variable displays the highest value
isei_10 - isei_77	indicator of respondents' International Socio-Economic Index (ISEI) value at ages 10 to 77 (based on ISCO-08). In case of multiple jobs, the variable displays the highest value
profstatus_10 - profstatus_77	indicator of respondents' professional status at ages 10 to 77
pubserv_10 - pubserv_77	indicator whether the respondent is employed in the private or public sector at ages 10 to 77
fedwork_10 - fedwork_77	indicator of the federal state of the work place at ages 10 to 77
emppara_10 - emppara_77	indicator of respondents' average number of parallel monthly employment spells at ages 10 to 77
empmonths_10 - empmonths_77	indicator of respondents' number of months in employment at ages 10 to 77
workh_c_10 - workh_c_77	indicator of respondents' average contractual weekly working hours at ages 10 to 77. Note that, unlike the variable workhours, this variable has not been plausibility checked and edited
workh_a_10 - workh_a_77	indicator of respondents' average actual weekly working hours at ages 10 to 77. Note that, unlike the variable workhours, this variable has not been plausibility checked and edited

Table 15, continued

Age-specific information from the prospective panel component	
jobs_21 - jobs_76	indicator of respondents' number of parallel jobs at ages 21 to 77
gross_earn_21 - gross_earn_76	indicator of respondents' gross monthly earnings from gainful employment at ages 21 to 76
net_earn_21 - net_earn_76	indicator of respondents' net monthly earnings from gainful employment at ages 21 to 76
gross_bonus_21 - gross_bonus_76	indicator of respondents' annual bonus payments from gainful employment at ages 21 to 76
gross_bonus_miss_21 - gross_bonus_miss_76	binary indicator whether respondents' information on annual bonus payments at ages 21 to 76 contains missing information
workhours_21 - workhours_76	indicator of respondents' actual total weekly working hours at ages 21 to 76

References

Bittmann, F. (2024). *plnepsmiss: Stata module run nepsmiss in parallel* [Computer software]. https://github.com/fbittmann/plnepsmiss

Blossfeld, H.-P., & Rossbach, H.-G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2. Auflage). Springer VS.