NEPS
National Educational Panel Study

Research Data

FDZ-LIfBi

Data Manual

NEPS Starting Cohort 3—Grade 5

Paths Through Lower Secondary School

Scientific Use File Version 12.0.0

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3, 96047 Bamberg
Director: Prof. Dr. Cordula Artelt
Administrative Director: Dr. Stefan Echinger
Bamberg; March 15, 2023

Research Data Documentation

The *NEPS Research Data Documentation Series* presents resources prepared to support the work with data from the National Educational Panel Study (NEPS).

Full citation of this document:
FDZ-LIfBi. (2023). *Data Manual NEPS Starting Cohort 3–Grade 5, Paths Through Lower Secondary School, Scientific Use File Version 12.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

This data manual for Starting Cohort 3–Grade 5 "Paths Through Lower Secondary School" has been prepared by the staff of the Research Data Center at the Leibniz Institute for Educational Trajectories (FDZ-LIfBi). It represents a major collaborative effort.

The contribution of the following persons is gratefully acknowledged:

Dietmar Angerer
Nadine Bachbauer
Daniel Fuß
Lydia Kleine
Tobias Koberg
Gregor Lampel
Sven Pelz
Benno Schönberger
Mihaela Tudose
Katja Vogel

Section 5 on Special Issues has been contributed by the following colleagues:

Agnieszka Althaber, Alexander Dicks, Teresa S. Friedrich, Cindy Fitzner, Insa Grüttgen, Alexander Helbig, Marie-Christine Laible, Josefine C. Matysiak, Laura Menze, Juliane Pehla, Ralf Künster, Benjamin Schulz, Annette Trahms, Basha Vicari

We also appreciate the work of former colleagues at the Research Data Center:

Daniel Bela, Simon Dickopf, Thomas Leopold, Manuel Munz, Sebastian Pink, Marcel Raab, Jan Skopek, Knut Wenzig, Markus Zielonka

Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe, LIfBi)
Research Data Center (Forschungsdatenzentrum, FDZ)
Wilhelmsplatz 3
96047 Bamberg, Germany

E-mail: fdz@lifbi.de
Web: https://www.lifbi.de/Institute/Organization/Research-Data-Center
Phone: +49 951 863 3511

Contents

1 Introduction

1.1 About this manual

This manual is intended to facilitate your work with data of NEPS Starting Cohort 3–Grade 5 (NEPS SC3). It serves both as a first guide for getting started with the complex data and as a reference book. The primary emphasis is on practical aspects such as sample development, data structure, and variable merging. The manual is neither complete nor exhaustive, but several links to other resources are provided in the respective paragraphs.

The first chapter refers to further documentation material, requirements for data access, instructions for data citation, some general rules and recommendations, and selected user services. In the second chapter, the fundamental objectives of Starting Cohort 3 and its sampling strategy are briefly introduced. The main part of this chapter is devoted to the sample development across the waves including field times, realized case numbers, survey modes, and the measurement of competency domains. The principles of Scientific Use File data-editing processes as well as conventions for naming the data files and variables are explained in the third chapter, supplemented by missing value definitions and an overview of additionally generated variables. The fourth chapter focuses on the data structure with information about data types, identifiers, and short portraits of all available datasets in the Scientific Use File. These short portraits include recommendations on how to use the dataset as well as syntax examples for merging variables of this dataset with variables from other files. The last chapter addresses some specific issues that should be noted when working with data of Starting Cohort 3. In the manuals for Starting Cohort 3 and 6 this section provides very detailed explanations of how the biographical life history data were collected and how they are stored in the various spell datasets in the Scientific Use File.

According to the cumulative release strategy – each new Scientific Use File contains the data of all previous survey waves plus the data of the currently prepared wave(s) – this manual will be regularly updated and revised for ongoing starting cohorts. While the information provided remains valid over time, at least the sample development must be continuously updated.

1.2 Further documentation

The data manual cannot cover all issues of data documentation in detail. Hence, a bunch of supplementary reports and other materials with background information on data preparation, survey instruments, competence tests, and field work (see Figure 1) can be downloaded from our website:

→ www.neps-data.de > Data Center > Data and Documentation
> Starting Cohort Grade 5 > Documentation

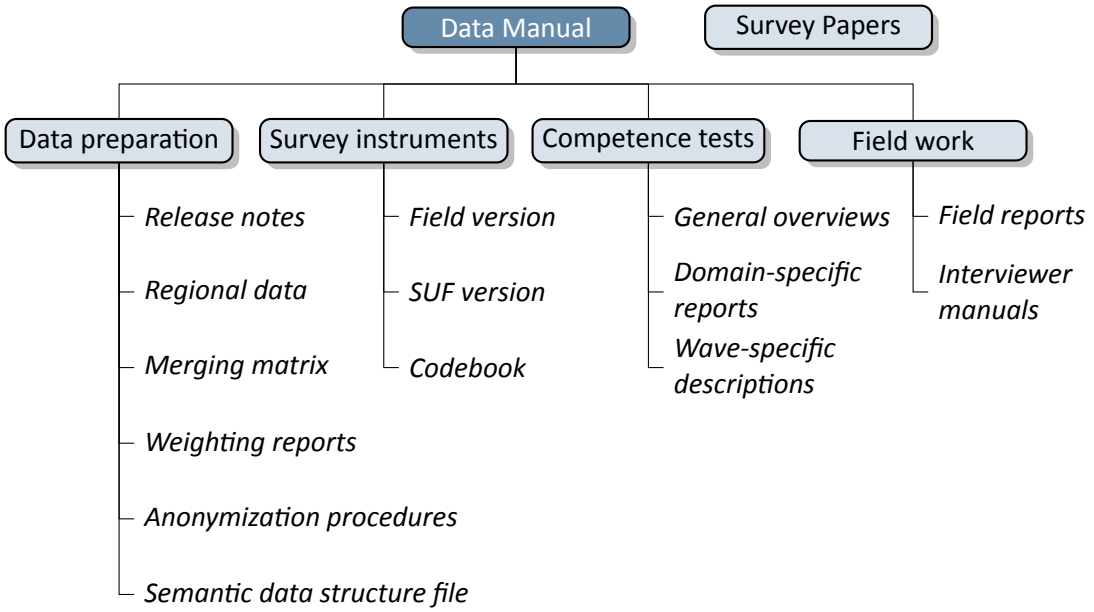

Figure 1: NEPS supplementary data documentation

Release notes All Scientific Use Files are accompanied by release notes that log changes in the data compared to prior versions and list bugs eliminated or at least known. For the latter, short syntax corrections are usually given. Please consult these notes when working with the data. See also section B.2 for a depiction of the current notes.

Regional data Fine-grained regional indicators from commercial providers (microm, RegioInfas) are available in our On-site environment. The report describes the regional levels covered by these indicators, their content, and how to merge them to the survey data.

Merging matrix This matrix provides an overview of how to link information from different datasets, taking into account the relevant identifier variables.

Weighting reports These reports entail information regarding the design principles of the sampling process and the creation of weights.

Anonymization procedures The document describes the anonymization measures carried out and provides an overview regarding the opportunity to access sensitive data.

Semantic data structure file This data package corresponds to the Scientific Use File but does not contain any observations (*purged datasets*). It provides all metadata including variable names, labels and answering scheme options to be used for exploring the data structure and for preparing analyses.

Survey instruments For each wave, the survey instruments are offered in the form of Scientific Use File (SUF) and field versions. While the field versions consist of the originally deployed instruments (in German only), the SUF versions are enriched by additional information

such as variable names and value labels used in the Scientific Use File. **Please note, that the competence test booklets are not publicly available**.

Codebook The codebook lists all variables and their corresponding labels plus the basic frequencies by waves in concordance with the datasets in the Scientific Use File.

Competence tests Information about competence testing is provided in various documentations, including general overviews and wave-specific descriptions; also, for each domain there is usually a brief description of the construct with sample items, a description of the data, and of the psychometric properties of the test.

Field reports The field reports document the overall data-collection process conducted by the survey institute. This information about survey preparation, interviewer deployment, respondent tracking, initial contacts, incentives, and sample realization is available in German only.

Interviewer manuals The interviewer manuals are a collection of instructions for the interviewers. In particular, they exemplify the interview process as well as the content of each of the questionnaire modules. They are available in German only.

NEPS Survey Papers Finally, there is a series of NEPS Survey Papers that address several topics of more general interest. These papers are listed for download on our website at:

→ www.neps-data.de > Data Center > Publications > NEPS Survey Papers

Additional documentation material might be available for this NEPS starting cohort. Please visit the data documentation website mentioned above for further details.

1.3 Data release strategy

NEPS data are published in the form of Scientific Use Files. They are provided free of charge to the scientific community. Each Scientific Use File consists of multiple datasets, forming a complex data structure with cross-sectional, panel and episode or spell information (see section 4). The release of NEPS Scientific Use Files follows a cumulative strategy, i. e., the latest data release replaces all former data releases. Hence, it is recommended to use the most current release of a Scientific Use File.

File Format

All Scientific Use Files are disseminated in Stata and SPSS format with bilingual variable labels and value labels in German and English. In the SPSS format, there are separate data files for both languages. Data stored in Stata format contain both languages within one file; the switch is induced by the following Stata command:

```
label language [de/en]
```

Due to the change of encoding to “Unicode” in Stata14 and the fact that older Stata versions are not able to open such data files, the NEPS Scientific Use Files contain two Stata formats, namely Stata14 and Stata(12-).

Versioning and Digital Object Identifier

Every time a new Scientific Use File is released, the data files existing up to now are either extended, usually by information from a new survey wave, or updated with changes due to larger or smaller corrections. The three digits of the version number inform about the number of waves integrated in the specific Scientific Use File, the frequency of major updates, and the frequency of minor updates. The version number is part of all relevant designations: that of the Scientific Use File, its data files (see Table 3), and the respective Digitial Object Identifier.

Every release of a NEPS Scientific Use File is registered at da|ra and clearly labeled with a unique *Digital Object Identifier* (DOI, see Wenzig, 2012). This DOI has two main functions: On the one hand, it enables researchers to cite the utilized NEPS data in an easy and precise way (see section 1.5). This in turn is a basic precondition for any replication analysis. On the other hand, the DOI directs to a landing page with further information about the Scientific Use File and the data access options. The DOI of the current release is doi:10.5157/NEPS:SC3:12.0.0. Other releases of Scientific Use Files for Starting Cohort 3 can be accessed by substituting the version number at the end of the DOI and the URL respectively (see Table 1).

Table 1: Release history of Scientific Use Files in Starting Cohort 3

SUF Version	DOI	Date of release
12.0.0 (current)	doi:10.5157/NEPS:SC3:12.0.0	**2022-11-28**
11.0.1	doi:10.5157/NEPS:SC3:11.0.1	2021-12-09
11.0.0	doi:10.5157/NEPS:SC3:11.0.0	2021-11-22
10.0.0	doi:10.5157/NEPS:SC3:10.0.0	2020-10-02
9.0.0	doi:10.5157/NEPS:SC3:9.0.0	2019-11-18
8.0.1	doi:10.5157/NEPS:SC3:8.0.1	2019-05-07
8.0.0	doi:10.5157/NEPS:SC3:8.0.0	2019-01-11
7.0.1	doi:10.5157/NEPS:SC3:7.0.1	2018-03-07
7.0.0	doi:10.5157/NEPS:SC3:7.0.0	2017-12-21
6.0.1	doi:10.5157/NEPS:SC3:6.0.1	2017-04-28
6.0.0	doi:10.5157/NEPS:SC3:6.0.0	2017-03-21
5.0.0	doi:10.5157/NEPS:SC3:5.0.0	2016-09-30
4.0.0	doi:10.5157/NEPS:SC3:4.0.0	2016-06-17
3.1.0	doi:10.5157/NEPS:SC3:3.1.0	2015-10-28

(...)

Table 1: (continued)

SUF Version	DOI	Date of release
3.0.0	doi:10.5157/NEPS:SC3:3.0.0	2015-08-31
2.0.0	doi:10.5157/NEPS:SC3:2.0.0	2014-01-31
1.0.0	doi:10.5157/NEPS:SC3:1.0.0	2012-10-01

1.4 Data access

Access to the NEPS data is free of charge but limited to the purpose of research and to members of the scientific community. Granting the right to obtain the data requires the conclusion of a Data Use Agreement. The existence of a valid Data Use Agreement entitles to work with all NEPS Scientific Use Files, i. e., the full data portfolio is at the disposal of the data recipient and all further persons involved in the agreement.

Application for data access

- Fill in the online form for a NEPS Data Use Agreement either in German or in English. Enter a title, the duration, and a short description of the intended research project. Make sure that all project participants with NEPS data access are specified in the form and that these persons have signed the agreement. Submit one copy of the complete agreement by e-mail, fax, or post. Further instructions and the relevant forms are provided on our website at:

  → www.neps-data.de > Data Center > Data Access > Data Use Agreements

- After approval by the Research Data Center, each registered NEPS data user receives an individual user name and a password to log in to our website. The basic Data Use Agreement permits the download of all available Scientific Use Files from our website at:

  → www.neps-data.de > Data Center > Data and Documentation > NEPS Data Portfolio

- There are two other modes of access to the NEPS data (see below); each demanding a supplemental agreement in addition to the basic Data Use Agreement.

- Another form is provided to state changes of the Data Use Agreement regarding further project participants or a prolonged project duration.

Modes of data access

Three modes of accessing the NEPS Scientific Use Files are available. They are designed to support the full range of researchers' interests regarding data utility while complying with the national and international standards of confidentiality protection. Each modus corresponds to a Scientific Use File version that is different in terms of accessibility of sensitive information.

- *Download* from the website = highest level of anonymization
- *RemoteNEPS* as browser-based remote desktop access = medium level of anonymization
- *On-site* access at secure working stations at LIfBi = lowest level of anonymization

While working with RemoteNEPS requires a biometrical authentication and Internet access, the On-site use of NEPS data necessitates a guest stay at LIfBi in Bamberg. More details about the access modes and their implications for application and utilization are provided at:

→ www.neps-data.de > Data Center > Data Access

Sensitive information

The download version of a Scientific Use File contains the least amount of information. For instance, institutional context data (xInstitution) or the Federal State label (*Bundesland-kennung*, see section 1.7) are only available in the controlled environments of RemoteNEPS and On-site. Indicators of a certain sensitivity are modified in the download data, such as aggregated categories for countries of citizenship or languages of origin. A few datasets and variables are exclusively accessible in the On-site version of a Scientific Use File, e. g., fine-grained regional indicators or open text entries. For more details see:

→ www.neps-data.de > Data Center > Data Access > Sensitive Information

This concept of nested data dissemination translates into an onion-shaped model of datasets. The most sensitive on-site level represents the outer layer with the remote and download levels being subsets of these data. That is, any data contained within a less sensitive level are included in the higher level(s). A detailed list of variables offered at the different levels together with notes on the generation of the three data versions can be found for each release of a Scientific Use File in the respective report on “Anonymization Procedures”.

1.5 Publications with NEPS data

Referencing the use of data from the National Educational Panel Study (NEPS) is essential for a good scientific practice as well as for revealing the scientific value of this study. The following citation rules apply to all publications based on NEPS data of Starting Cohort 3.

It is obligatory to acknowledge the NEPS study in general and to indicate the utilized data version by citing the data version (DOI) as follows:

> NEPS Network. (2022). *National Educational Panel Study, Scientific Use File of Starting Cohort Grade 5*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC3:12.0.0

In addition, the NEPS study is to be referred to at an appropriate place:

> This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld and Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network.

Finally, the reference article should be listed in the bibliography:

> Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition ZfE* (2nd ed.). Springer VS. https://doi.org/10.1007/978-3-658-23162-0

Authors of any kind of publications based on NEPS data are requested to notify the Research Data Center about their articles and to provide an electronic version or a copy. All reported publications are listed in the NEPS Bibliography on our website at:

→ www.neps-data.de > Data Center > Publications

Citing documentation

To refer to any of the documentation material published in the *NEPS Research Data Documentation Series* (e. g., this manual), please make use of the following citation templates:

> FDZ-LIfBi. (2023). *Data Manual NEPS Starting Cohort 3–Grade 5, Paths Through Lower Secondary School, Scientific Use File Version 12.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Or another example:

> Schönberger, K., & Koberg, T. (2017). *Regional Data: Microm*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

If no author is given, please take a universal *NEPS Network* instead:

> NEPS Network. (2023). *Starting Cohort 3: Grade 5 (SC3), Wave 12, Questionnaires (SUF Version 12.0.0)*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

If a document has not been published in this series, please refer to the author and the title as in the following citation of a field report by one of the survey institutes:

> Kersting, A., & Aust, F. (2019). *Methodenbericht. NEPS Startkohorte 3 (Schulabgänger und individuell nachverfolgte Schüler) – Haupterhebung Herbst 2018, Teilstudie B132*. Bonn, Germany: infas Institut für angewandte Sozialwissenschaft GmbH.

1.6 Rules and recommendations

Working with NEPS data is bound to a couple of rules that are codified in the Data Use Agreement. Each data user has to confirm these rules by his or her signature. The already mentioned obligation to cite the study and to indicate any kind of publication resulting from the use of NEPS data (see section 1.5) are just two examples. The major part of rules refers to issues of data privacy and the requirements of careful data handling.

Rules

- *Avoidance of re-identification:* Any action aimed at and suitable for re-identifying persons, households, or institutions is strictly forbidden. This also includes the combination of NEPS data with other data that allow for a re-identification of persons. In case of any accidental re-identification, the Research Data Center has to be informed immediately and all individual data gained therefrom have to be kept secret.
- *Avoidance of data disclosure:* NEPS data are exclusively provided on the basis of a valid Data Use Agreement – for a defined purpose (research project) and to a defined group of persons (data recipient and further project members that are involved in the contract). Any use for commercial or other economic purposes is not permitted just as any transfer of the data to third parties. Please handle the provided NEPS data with strict confidentiality!
- *Regulations on using the Federal State label:* For NEPS data collected in connection with schools or higher education institutions it is not allowed to use Federal-State-related information directly or indirectly contained in the data for analyses aiming at direct comparisons of the German Federal States (*Bundesländer*), or aiming at direct conclusions to be drawn about a Federal State, or aiming at a reconstruction of the concrete Federal State affiliation of persons, households, and institutions. Any kind of ranking between the Federal States based on NEPS data is prohibited (see section 1.7).

Please note that violation of these rules may lead to severe penalties as stated in the NEPS Data Use Agreement. If there is any doubt or question regarding the given regulations, please contact the Research Data Center (see section 1.9). The same applies in case of encountering any deficiencies in data quality or any security leaks with regard to NEPS data protection and data security.

Recommendations

In addition to the binding rules, there are some recommendations for the use of NEPS data:

- *As a matter of course:* Always be critical when working with empirical data. Although a big effort is being made to ensure the integrity of the provided data we cannot guarantee absolute correctness. Notices on problems or errors in the datasets are welcome at any time at the Research Data Center.
- *Enhanced understanding of the data:* Consult the documentation and survey instruments. The analysis of complex data necessitates a precise idea of how the information were collected and edited. All relevant material is available online (see section 1.2).
- *Facilitated handling of the data:* Utilize the tools that are offered. Several user services are provided to support NEPS data analyses – reaching from specific Stata commands (e. g., for an easy recoding of missing values) to a meta search engine (e. g., for an interactive exploration of all instruments) to an online discussion forum (e. g., for the clarification of questions). These tools are also available online, see section 1.8 for more details.

1.7 On using the Federal State label *(Bundeslandkennung)*

In concurrence with the regulations of the Research Data Center at the Institute for Educational Quality Improvement (Institut zur Qualitätsentwicklung im Bildungswesen, IQB), using the Federal State label in conjunction with NEPS data collected in connection with schools or higher education institutions is permitted in the context of exploring scientific research questions, if it is exclusively used for:

- control purposes in order to incorporate it as a covariate in the planned analysis; the identification of individual Federal States in the displayed results is not permitted
- incorporating contextual characteristics or other third-party variables; the identification of individual Federal States in the displayed results is not permitted
- comparing aggregated groups of Federal States where at least two states are combined to form a single meaningful group with regard to substantive issues; the identification of individual Federal States in the displayed results is not permitted
- for sample descriptions (e. g., the distribution of participants by state and by different types of schools within states)

When using data collected in connection with schools or higher education institutions, it is **not allowed** to use Federal-State-related information directly or indirectly contained in the data for analyses aiming at a direct Federal State comparison, direct conclusions to be drawn about a Federal State, or a reconstruction of the concrete Federal State affiliation of persons, households, and institutions.

The Federal State label in the starting cohorts of schools and higher education institutions is provided to the scientific community only via remote access (*RemoteNEPS*) and – depending on availability – via guest working stations in Bamberg (*On-site*). The respective analysis results are reviewed by staff of the Research Data Center before being passed on electronically to the researcher in a password-protected environment. The abovementioned restrictions concerning the use of the Federal State label do not apply to data collected in a nonschool context and/or in Federal-State-specific educational reform studies.

1.8 User services

In addition to a comprehensive data documentation, there are several user services to support researchers working with the NEPS data. First and foremost, the Research Data Center maintains a regularly updated and enhanced website with detailed information on all Scientific Use Files, a complete list of NEPS projects, a NEPS bibliography, a reference to NEPS events, and a NEPS newsletter. All subsequently introduced services and tools can be reached via this website:

→ www.neps-data.de > NEPS

Forum4MICA

The *Forum4MICA – Making Information Commonly Available* is an open online discussion platform for experienced users as well as for persons who are just searching for relevant information. The forum is joined by various Research Data Centers with their data collections, including the FDZ-LIfBi with the NEPS data. It offers the opportunity to exchange with NEPS staff members and with other researchers in a transparent dialogue. In this way, the forum grows into a knowledge archive with practical solutions to numerous problems and questions. We highly encourage you to browse it first when struggling with NEPS issues or when help is needed with specific data matters. If there is no solution available, please take the opportunity to share your question by posting it to the forum. Active participation requires no more than a one-time registration. The entire NEPS user community (and beyond) will benefit from a broad participation. You can find the *Forum4MICA* at:

→ `https://forum.lifbi.de`

NEPSplorer

The *NEPSplorer* facilitates an interactive and quick full text search through all instruments of released NEPS surveys, with the exception of competence tests. The tool is particularly suitable for getting a first idea of the availability of constructs, items, and variables in the datasets. It is based on both keyword search with several filtering options and hierarchical construct search. The *NEPSplorer* offers some helpful functions such as displaying univariate statistics, listing relevant metadata, and enabling registered users to create their own personal watch list of interesting items. As a web application – a mobile version aligned for smartphone usage is also available – the *NEPSplorer* relies on the most up-to-date information; any correction in the metadata is thus instantly visible. Start the tool here:

→ `www.neps-data.de` > `Data Center` > `Overview and Assistance` > `NEPSplorer`

NEPStools

NEPStools is a free to use collection of Stata commands that is created and supplied by the Research Data Center. The package includes some programs (“ado files”) that make NEPS data handling easier. As an example, the `nepsmiss` command automatically recodes all of the numeric missing values (-97, -98, etc.) into Stata’s “Extended Missings” (.a, .b, etc.) with correctly recoded value labels. Another example ist the `infoquery` command that displays additional attributes of the variable such as the question text and the initial variable name in the instrument. *NEPStools* can be installed from our repository through Stata’s built-in installation mechanism:

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

A description of the programs and further information are given on the website at:

→ `www.neps-data.de` > `Data Center` > `Overview and Assistance` > `Stata Tools`

NEPSscaling

Plausible Values are a way of describing the competencies of individuals at the group level. They allow (unbiased) estimates of effects at the population level that are adjusted for measurement errors. In contrast to point estimators such as Weighted Likelihood Estimates (WLE), the use of Plausible Values is suitable for more precise inferential statistical tests in correlation and mean value analyses. The R package *NEPSscaling* enables users to generate own Plausible Values with a background model adapted to the specific research question. The package is able to handle missing values in the background model and has additional features. More information is available here:

→ www.neps-data.de > Data Center > Overview and Assistance > NEPSscaling

Data trainings

The Research Data Center offers a series of regular NEPS data trainings, usually conducted as online courses. Participation in the one- or two-day courses is free of charge. The courses consist of different modules, whereby single modules can be attended separately. While the *basic modules* provide knowledge on the general framework of the NEPS study and on how to access and work with the NEPS data plus documentation, the *advanced modules* address selected topics such as the handling of competence data, episode data, linked NEPS-ADIAB data, weights, etc. A schedule of current training courses together with information for registration can be found at our website:

→ www.neps-data.de > Data Center > Data Trainings

1.9 Contacting the Research Data Center

The Research Data Center at the Leibniz Institute for Educational Trajectories (Forschungsdatenzentrum, FDZ-LIfBi) accounts for large parts of the NEPS data preparation and documentation, the data dissemination, and the user support including individual advice. We welcome your feedback to further improve our products and services. This particularly applies to this manual as the guiding document to facilitate your work with NEPS data of Starting Cohort 3.

Please contact us with your questions, comments, requests, and suggestions:

E-mail: fdz@lifbi.de
Web: → www.neps-data.de > Data Center > Contact Data Center
Phone: +49 951 863 3511

2 Sampling and Survey Overview

2.1 Paths through lower secondary school

The lower secondary level plays a connecting role between elementary school and the general or vocational upper secondary level (or direct entry into the job market). However, important questions have not yet been answered clearly and conclusively due to the lack of appropriate data. This concerns, for example, the type of school chosen, the switch to another type of school, or the repetition of class levels, but also the central issue of paths through lower secondary level and the transition to upper secondary level.

The study *Paths Through Lower Secondary School: Educational Pathways of Students in Grade 5 and Higher* by the National Educational Panel Study (NEPS) follows a sample of representatively selected students who attended regular or special schools and were willing to participate in annual surveys and competency tests. After leaving school or the general school system, these students are further interviewed and individually tested outside the educational institution.

Key questions of this study relate to the development of students' competences as well as to the conditions and prerequisites of their educational processes. In addition, the focus is on possible personal consequences for success and future courses of education as well as the integration of students into social networks. The questionnaires for the teaching staff and the school principals of the participating schools collect data, for example, on class size, the composition of the student body and school equipment, but also contain questions on teaching in general.

2.2 Sampling strategy

The target population of the Starting Cohort 3 includes students of the 5th grade in German schools, which offered lower secondary education in the school year 2010/11. Access to this population of children was through the schools they attended. The drawing of the initial sample took place in a stratified two-stage procedure, in which first schools and then classes within the schools were drawn at random. For the NEPS starting cohorts 3 and 4, school sampling was carried out in one common step, i.e., schools were drawn for the surveys in grades 5 and 9 at the same time.

On the first stage, a complete list of all regular schools in Germany as primary sampling units (PSU) formed the basis for the selection. This list contained only schools of the general education system (e. g., no vocational schools). It was compiled with the help of up-to-date school directories for the school year 2008/2009, as they were available to the *Statistical Offices* of

all 16 federal states. In order to adequately reflect the diversity of the German federal state-specific school systems, all schools on the list were classified according to their type and the classification used as one out of four characteristics for implicit stratification:

- school type (*Gymnasien, Realschulen, Hauptschulen, Schulen mit mehreren Bildungsgängen, Gesamtschulen, Grundschulen, Förderschulen*)
- federal state
- regional classification
- sponsorship (public vs private)

From the total list of 11,792 schools (excluding special-needs schools), 240 original schools were selected for grade 5. 214 of these 240 schools were also included in the initial sample for the 9th grade survey of Starting Cohort 4. In addition, four substitute schools were drawn for each selected school – with identical characteristics with regard to the federal state, sponsorship, regional classification and school type as well as similar class sizes. The school sample was supplemented by 65 special-needs schools with a focus on "learning" selected from the list and three substitute schools each. All drawn special-needs schools overlap with the sample of Starting Cohort 4.[1]

The participation in the study was voluntary for the selected schools as well as for the students. Despite the support of the *Ministries of Education* of the federal states, the recruitment of the schools proved to be a particular challenge. If an original school refused to take part in the NEPS, the loss was compensated by taking into account one of the substitute schools drawn in addition to this school.

On the second stage, two fifth classes were drawn at random in each participating school. In these two classes, all students as secondary sampling units (SSU) were asked to take part in the panel study. Prior to this selection step, the recruited schools were contacted by the survey institute and asked for information on all classes in grade 5, including the number of students per class. If there were more than two classes in a school, the two classes were selected by a "systematic random start interval sampling" procedure. If there was only one fifth class in a school, only that class was selected. In the special-needs schools, a simple random sampling was carried out, i.e., all fifth-graders of the participating schools were asked to participate.

In order to be part of the study, parents had to give their written consent. Only children for whom a fully completed consent form was available on the day of the survey were allowed to take part. The individual forms were distributed to the students via the schools and collected again there. Of the total of 9,622 reported students in the selected fifth classes at regular schools (gross sample), altogether 5,283 students were willing to participate in the NEPS study. These children were in possession of the mandatory declarations of consent from their

1 A further sample of students with a Turkish migration background or a migrant background from the former Soviet Union supplements the cohort sample of fifth-graders. The basis for this additional study was the drawing of 57 original schools (and 30 substitute schools) with a particularly high proportion of students with a corresponding migrant background.

parents.[2] 4,989 out of these 5,283 target persons took the tests in the first survey wave and/or completed the questionnaire. This corresponds to a participation rate of 94.4 percent. From the selected special-needs schools, a total of 1,064 students were reported. Parental consent was obtained for 587 students, and 566 students finally took part in the survey. Accordingly, the participation rate for this subsample is 97.1 percent. In sum, the first wave of the fifth-grade sample contains information of 5,774 students from 291 schools.[3]

Refreshment sample

In some federal states, namely Berlin and Brandenburg, the transition to lower secondary education takes place after grade 6. Accordingly, when students change schools after completing grade 6, they also leave the institutional context in which they were originally sampled and interviewed or tested. Against this background, a refreshment sample of seventh-graders was established in the course of the third survey wave in order to compensate for this loss of students in their institutional context.

The drawing of the refreshment sample was largely analogous to the drawing of the main sample of regular schools. Only schools that had not yet been included in the first draw for the starting cohorts 3 or 4 were eligible for the refreshment. Again, a two-stage sampling design with explicit and implicit stratification was applied. The two explicit strata reflect the different timings of the transition to lower secondary education. One stratum includes all regular schools in Berlin and Brandenburg, which have at least one grade 7, but do not offer grades 5 and 6. The second stratum contains all regular schools of the other 14 federal states with at least one grade 7. Within these two strata, altogether 100 original schools (and about 400 substitute schools) were systematically selected using probability proportional to size sampling. In the refreshment sample, 86 out of 374 contacted schools decided to participate in the study. On the second stage, two classes were randomly selected within these schools. If there were less than three seventh classes in a school, all classes of this grade were selected. All students from the selected classes were invited to participate.

A total of 3,944 seventh-graders were reported by the participating schools (gross sample). Of these, 2,205 were willing to take part in the NEPS study, i.e., they provided valid declarations of consent from their parents. In the end, altogether 2,146 children from the refreshment sample were surveyed in the third wave of Starting Cohort 3, i.e., their first measurement.

The sampling design and its consequences for the derivation of sampling weights are fully described in Steinhauer and Zinn, 2016a. Further remarks on the recruiting process are given in the PAPI field report of the first survey wave in regular as well as in special-needs schools (in German only). All documents are available on our website at:

→ www.neps-data.de > Data Center > Data and Documenation
> Starting Cohort Grade 5 > Documentation

2 In later waves, it became a requirement in some federal states that the students (at the earliest when they reached the age of 14) had to give their own consent in order to continue participating in the study.
3 From the additional sample, 219 fifth-graders with a Turkish migration background or a migrant background from the former Soviet Union took part in the survey of the first wave of Starting Cohort 3.

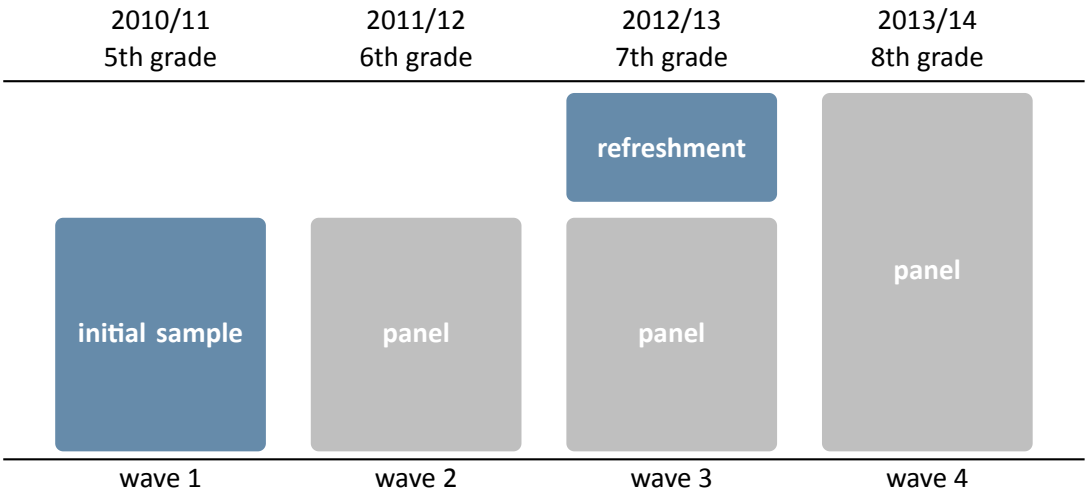

Figure 2: Longitudinal sampling design of Starting Cohort 3

Individual tracking

In addition to students who continue to attend school, those respondents who have left the general education system were surveyed separately. The so-called *individual tracking* or *individual follow-up* means that the data collection is no longer conducted in the institutional context of a school or class, but in the form of out-of-school surveys. The tracking began in fall 2015 with school-leavers after grade 9. Individual tracking also includes students for whom the NEPS school refused further participation in the study (dropout) as well as students who have left the original NEPS school but continued to attend another general education school (school transfer).[4] The interviews in the individual tracking took place in a similar time corridor as the parallel surveys at schools. However, the survey mode differed, as did parts of the survey program and competence testing. Individually followed-up respondents can be identified in the sample via the variable `tx80230` ("Panel Frame") in the dataset `CohortProfile`.

Context persons

Target persons of Starting Cohort 3 are students, beginning with the first survey in grade 5 or in the case of the refreshment sample in grade 7. In order to also collect information about the institutional learning environment, supplementary contextual data were surveyed from the class teachers, the German and mathematics teachers of participating classes as well as from the school principals. These additional surveys were conducted using paper questionnaires (PAPI) to be completed separately from the students' survey. Participation was also voluntary for these

4 Until fall 2015, target persons who changed schools within the general education system or where the school no longer participated in the NEPS study (due to cancellation of willingness to participate, school closure, discontinuation of the grade, or too low willingness to participate at the student level) were individually tracked by the survey institute IEA DPC–IEA Data Processing and Research Center, Hamburg. Starting in fall 2015, infas–Institute for Applied Social Sciences, Bonn, took over the follow-up of these students together with the group of school-leavers.

context persons. *Teachers* answered a person-related general questionnaire section as well as one or more subject-specific sections in the survey waves 1 through 5 and 7. *School principals* answered a questionnaire for school-related information in the survey waves 1 through 5 and 7 through 9. Please note, that the dataset with data from the principals (`xInstitution`) is only available in the RemoteNEPS and On-Site version of the Scientific Use File.

The family or home context of the target persons was captured by repeated interviews with one legal guardian. These *parent* interviews, which were also designed as a panel survey, took place in the waves 1 through 4 and 6. They were conducted in the form of computer-assisted telephone interviews (CATI). Whenever possible, the biological or social parent who was most familiar with the child's school matters was interviewed. In the vast majority of cases, this was the biological mother. During the study, it was possible to change to another parent with parental authority. Participation in the interviews was also voluntary; the corresponding consent was obtained separately from the willingness regarding the child's participation in the study. In terms of content, the parent surveys focused primarily on educational aspects and the school history of their children, but also on parental support for the children, the children's health, satisfaction with school, the use of language in the family, the household equipment, as well as on various sociodemographic characteristics, etc.

A detailed picture of the survey units, the realized case numbers, the survey modes and the responsible survey institutes for each survey wave is provided in section 2.4.

2.3 Competence measures

The collection and provision of data on the development of competencies and skills throughout the life course is a key element of the NEPS. Competence measurements are carried out across different waves in all NEPS starting cohorts covering *domain-general* and *domain-specific cognitive competencies* as well as *metacompetencies* and *stage-specific competencies*.

Data from the competence tests pass through an editing process before they get integrated into the Scientific Use File. This data preparation enables users to work with scored items and generated test scores such as the sum or mean of correct answers. Detailed descriptions on how these scores were estimated can be found in separate reports for the respective competence domains (see section 1.2). The scores are compiled in a dataset named `xTargetCompetencies`. This dataset is structured in the so-called wide format, that is, all responses of a single respondent are placed in one row of the data matrix.[5] As a consequence, variable names for competence scores follow a specific nomenclature. These conventions not only allow for the identification of the respective domain, the target group, the testing modus, and the kind of scoring, they also inform about the repeated administration of a test item in a different wave or starting cohort (see section 3.2.2).

5 The Scientific Use File contains another competence dataset (`xPlausibleValues`) with generated variables for plausible values (see Scharl and Zink, 2022 and section 1.8).

The next table shows the schedule of competence measures in Starting Cohort 3 with domains by waves and test modus.

- Subsequent to several competence tests (re, vo, li, ma, sc, nr/nt, ic, or, ef), the target persons had to assess their own test performance ("Procedural Metacognition", mp).[6]
- The L1-Test for Russian and Turkish language has been applied to target persons of a corresponding migration background only.
- Reduced testing: In wave 9, both stage-specific competence tests (ef, st) were realized in the institutional context only (without individually tracked target persons). For individually tracked target persons, a randomized allocation of competence tests with two out of the three domains (re + ma OR re + ic OR ma + ic) has been applied.
- The administration of the ICT-Literacy test in wave 9 was paper-based in the institutional context and computer-based for individually tracked target persons.
- In the survey wave 4 as well as from wave 10 onwards, there were no competence tests at all administered to the target persons.

6 The list of all possible competence domains together with the respective abbreviation can be found in table 5.

Table 2: Schedule of competence measures. P = Paper-Based Test (proctored), C = Computer-Based Test (proctored), LC = Listening Comprehension

		2010/11 **Wave 1** Grade 5	2011/12 **Wave 2** Grade 6	2012/13 **Wave 3** Grade 7	2014/15 **Wave 5** Grade 9	2015 **Wave 6** Grade 9	2015/16 **Wave 7** Grade 10	2016/17 **Wave 8** Grade 11	2017/18 **Wave 9** Grade 12
Domain-General Competencies									
DGCF: Cognitive Basic Skills	dg	P	—	—	—	P	—	—	—
Domain-Specific Competencies									
Reading Competence	re	P	—	P	—	P	—	—	P
Reading Speed	rs	P	—	—	P	—	—	—	—
Vocabulary: LC at Word Level	vo	—	P	—	—	—	—	—	—
Listening: LC at Text Level	li	—	—	—	—	P	—	—	—
Mathematical Competence	ma	P	—	P	P	—	—	—	P
Scientific Competence	sc	—	P	—	P	—	—	P	—
Native Language Russian/Turkish: LC	nr/nt	—	—	P	—	P	—	—	—
Metacompetencies									
Declarative Metacognition	md	—	P	—	—	P	—	—	—
ICT Literacy	ic	—	P	—	P	—	—	—	P/C
Stage-Specific Competencies									
Orthography	or	P	—	P	P	—	—	—	—
English Reading Competence	ef	—	—	—	—	—	P	—	P
Scientific Thinking	st	—	—	—	—	—	—	—	P

2.4 Survey overview and sample development

This section informs about the progress of the Starting Cohort 3 sample. For each survey wave in the current Scientific Use File, there is a short characterization in terms of field time, number of realized cases, relevant subsamples, survey modes, and the survey institute(s) responsible for collecting the data. A more detailed insight into all aspects of the field work is provided by the *Field Reports*, which are available on the website (in German only) as part of the data documentation for each NEPS (sub-)study.

→ www.neps-data.de > Data Center > Data and Documentation
> Starting Cohort Grade 5 > Documentation

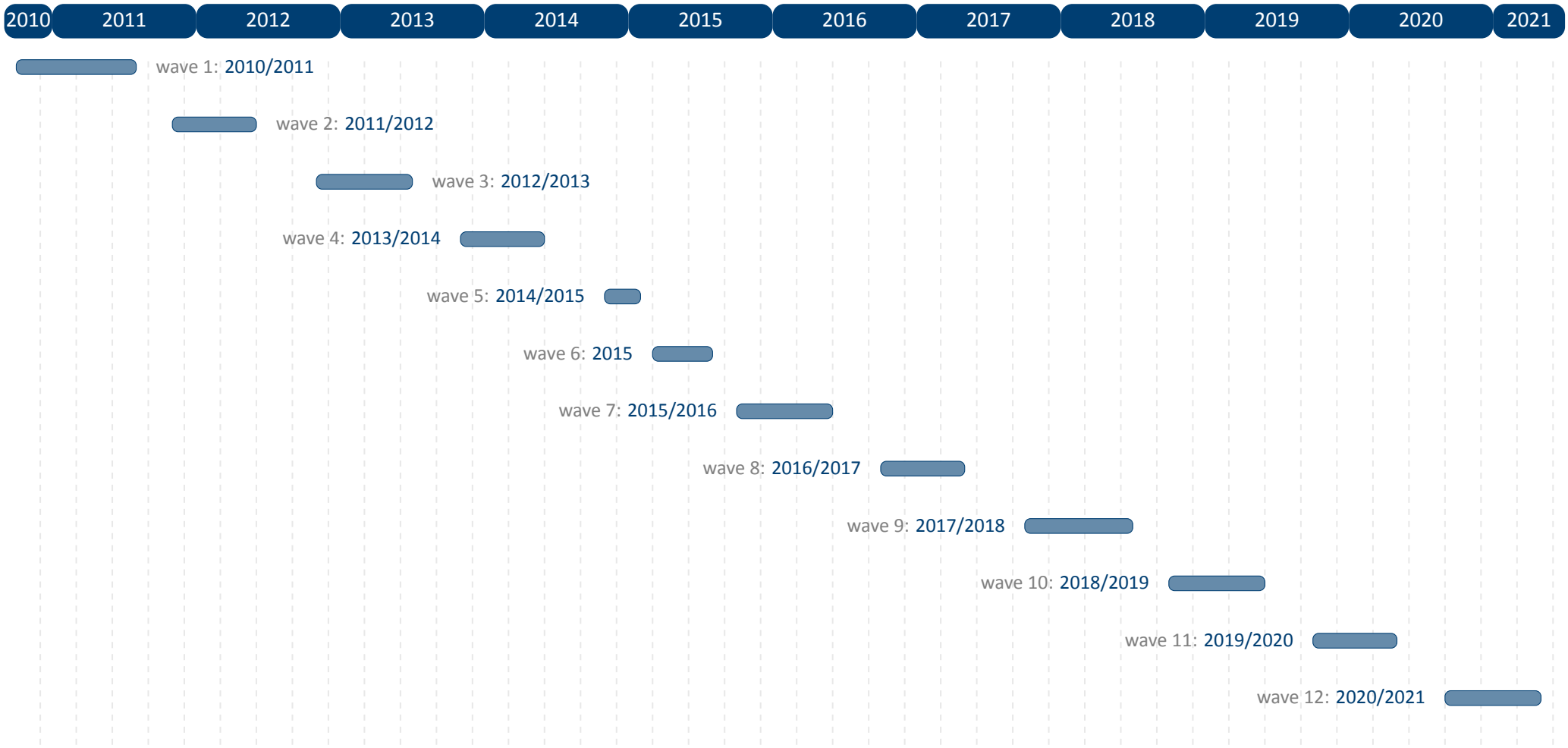

Figure 3: Panel progress of Starting Cohort 3 (waves 1 to 12)

2.4.1 Wave 1: 2010/2011

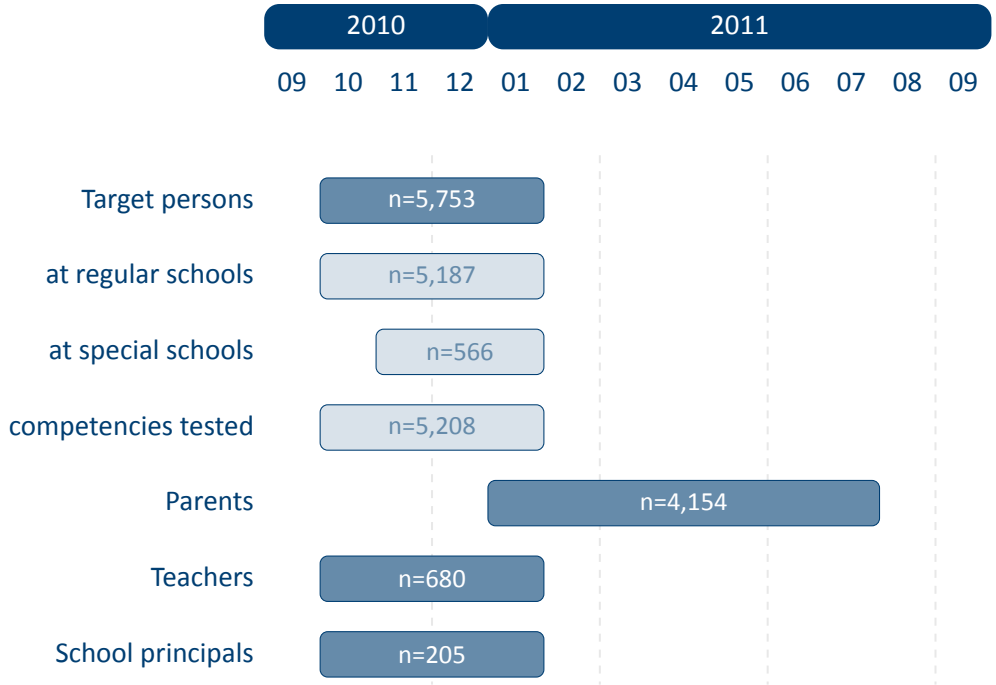

Figure 4: Field times and realized case numbers in wave 1

- **Target persons** 5th graders at panel start 2010/11
  - *Grade 5 students at regular schools*

    Sample Multi-stage stratified cluster sampling with the following selection stages:

    1. Random sampling of regular schools at lower secondary level. During this stage, the regular schools used in Starting Cohort 4–Grade 9 are redrawn. A size-proportional random selection is applied to select the regular schools for grade 5 from this pool. In addition, schools that teach students in grade 5, but not in grade 9 are included.
    2. Random selection of grade 5 classes within the selected schools. Two grade 5 classes are selected per school.
    3. All students of the selected classes are invited to participate in the study.

    Modus Written questionnaires and competence tests completed in class context (PAPI)

- *Grade 5 students at special schools*

  Sample Multi-stage cluster sampling with the following selection stages:

  1. Random sampling of special schools at lower secondary level. During this stage, the special schools used in Starting Cohort 4–Grade 9 are redrawn. A simple random selection is applied to select the special schools for grade 5 from this pool.
  2. All students attending grade 5 at the selected schools are invited to participate (=full sample survey of grade 5).

  Modus Written questionnaires and competence tests completed in class context (PAPI)

- *Individually tracked students*

  Size No survey in the individual tracking in this wave

- **Context persons**
  - *Parents*

    Sample One biological or social parent per target child

    Modus Computer-assisted telephone interviews (CATI)

  - *Teachers*

    Sample Class teachers of the target children and their teachers for the subjects German and mathematics

    Modus Written questionnaires (PAPI)

  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)

- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.2 Wave 2: 2011/2012

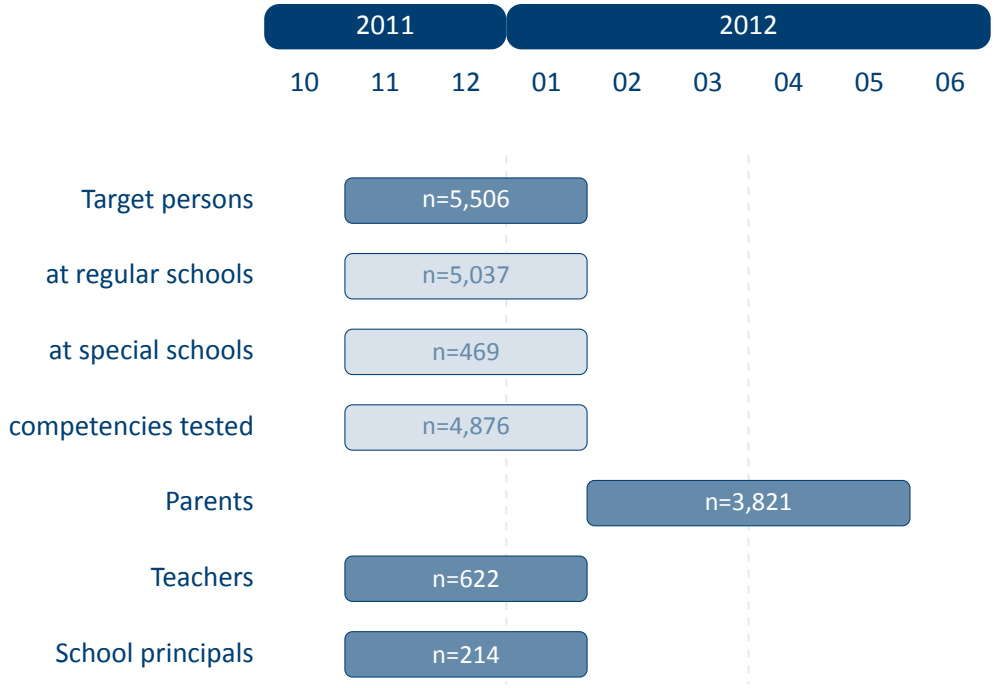

Figure 5: Field times and realized case numbers in wave 2

- **Target persons** 5th graders at panel start 2010/11
  - *Grade 6 students at regular schools*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Grade 6 students at special schools*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked students*

    Size 185 students in the individual tracking with survey and/or competence data

- **Context persons**
  - *Parents*

    Sample One biological or social parent per target child (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews (CATI)
  - *Teachers*

    Sample Class teachers of the target children and their teachers for the subjects German and mathematics

    Modus Written questionnaires (PAPI)
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.3 Wave 3: 2012/2013

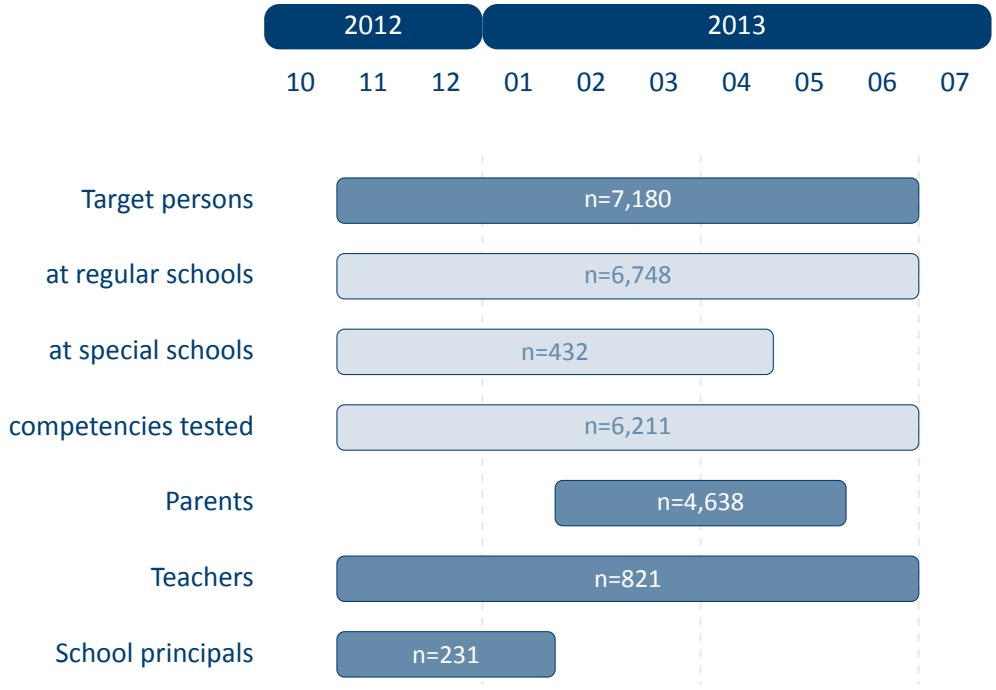

Figure 6: Field times and realized case numbers in wave 3

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 7 students at regular schools*

    Initial Sample Participants who continue to be willing to participate in the panel

    Refreshment Sample Students who participate for the first time in the survey and testing

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Grade 7 students at special schools*

    Initial Sample Participants who continue to be willing to participate in the panel

    Refreshment Sample Students who participate for the first time in the survey and testing

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked students*

    Size 568 students in the individual tracking with survey and/or competence data

- **Context persons**
  - *Parents*

    Sample One biological or social parent per target child (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews (CATI)
  - *Teachers*

    Sample Class teachers of the target children and their teachers for the subjects German and mathematics

    Modus Written questionnaires (PAPI)
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.4 Wave 4: 2013/2014

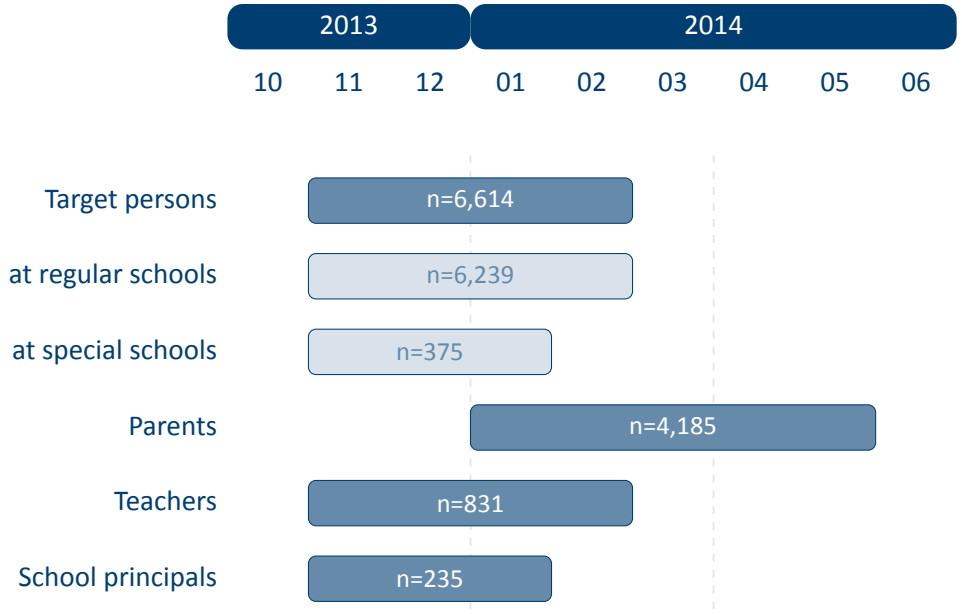

Figure 7: Field times and realized case numbers in wave 4

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 8 students at regular schools*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Grade 8 students at special schools*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked students*

    Size 675 students in the individual tracking with survey and/or competence data
- **Context persons**
  - *Parents*

    Sample One biological or social parent per target child (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews (CATI)

- *Teachers*

  Sample Class teachers of the target children and their teachers for the subjects German and mathematics

  Modus Written questionnaires (PAPI)

- *School principals*

  Sample Principals of all schools with participating classes

  Modus Written questionnaires (PAPI)

- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.5 Wave 5: 2014/2015

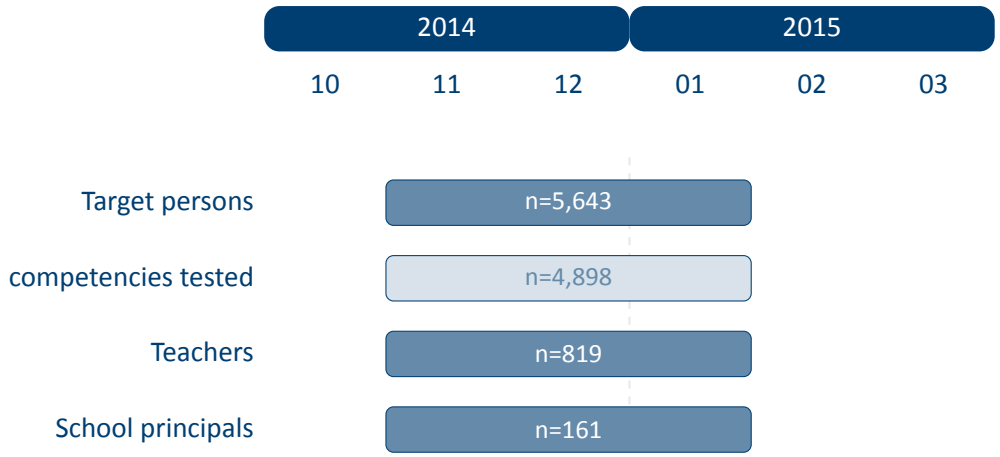

Figure 8: Field times and realized case numbers in wave 5

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 9 students*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked students*

    Size 769 students in the individual tracking with survey and/or competence data
- **Context persons**
  - *Teachers*

    Sample Class teachers of the target children and their teachers for the subjects German and mathematics

    Modus Written questionnaires (PAPI)
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.6 Wave 6: 2015

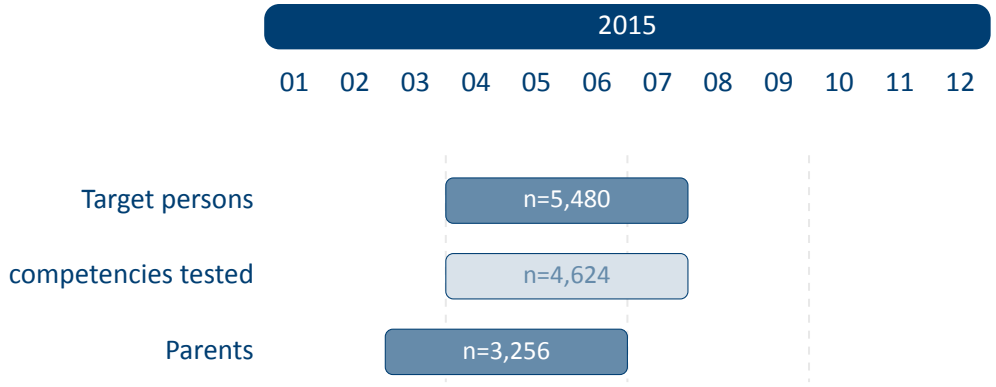

Figure 9: Field times and realized case numbers in wave 6

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 9 students*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked students*

    Size 886 students in the individual tracking with survey and/or competence data
- **Context persons**
  - *Parents*

    Sample One biological or social parent per target child (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews (CATI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Family context, CATI infas–Institute for Applied Social Sciences, Bonn

2.4.7 Wave 7: 2015/2016

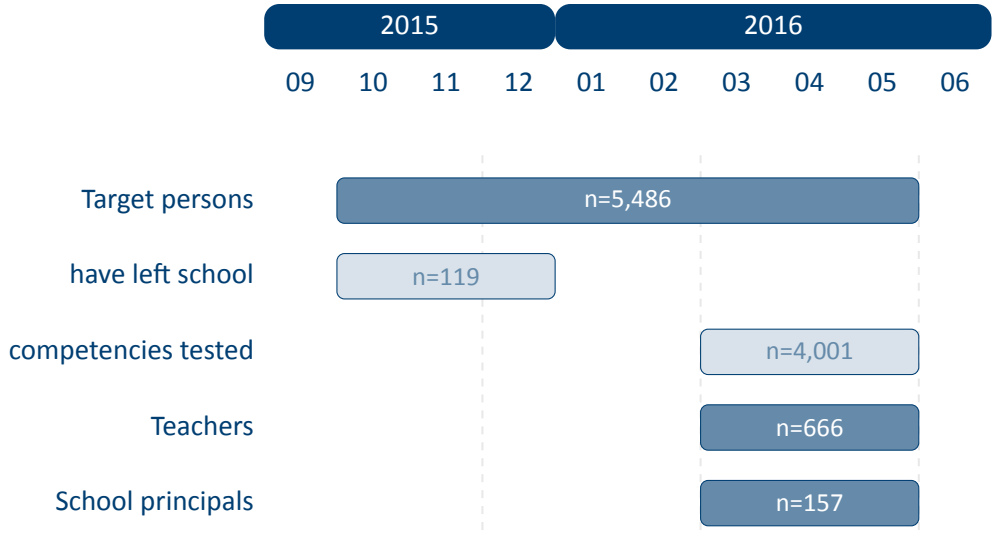

Figure 10: Field times and realized case numbers in wave 7

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 10 students*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted telephone interviews (CATI), no competence testing
  - *Individually tracked students*

    Size 1239 students in the individual tracking with survey data

    Modus Computer-assisted telephone interviews (CATI) and subsequent Online interviews (CAWI), no competence testing

- **Context persons**
  - *Teachers*

    Sample Class teachers of the target children and their teachers for the subjects German and mathematics

    Modus Written questionnaires (PAPI)
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Individual tracking, CATI/CAWI infas–Institute for Applied Social Sciences, Bonn

2.4.8 Wave 8: 2016/2017

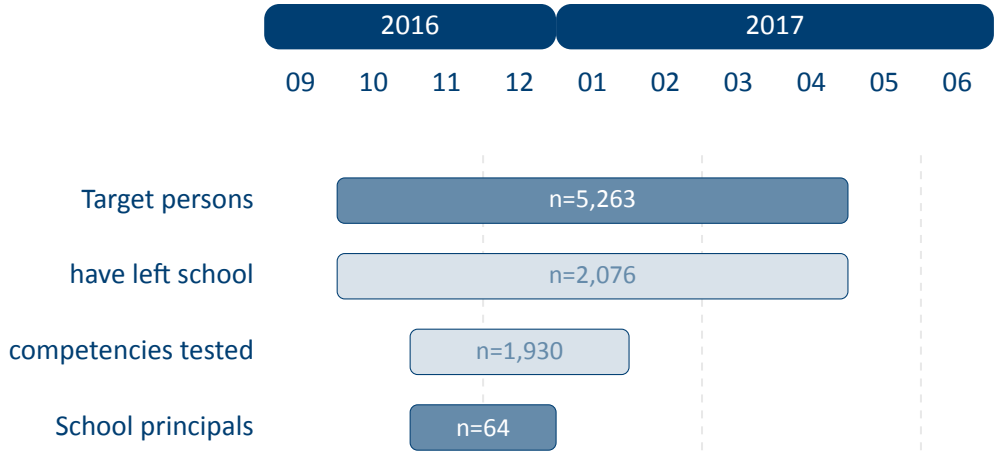

Figure 11: Field times and realized case numbers in wave 8

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 11 students*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted telephone interviews (CATI) or Computer-assisted personal interviews (CAPI) as alternative plus subsequent Online interviews (CAWI) for selected participants, no competence testing
  - *Individually tracked students*

    Size 1244 students in the individual tracking with survey data

    Modus Computer-assisted telephone interviews (CATI) or Computer-assisted personal interviews (CAPI) as alternative plus subsequent Online interviews (CAWI), no competence testing

- **Context persons**
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Individual tracking, CATI/CAPI/CAWI infas–Institute for Applied Social Sciences, Bonn

2.4.9 Wave 9: 2017/2018

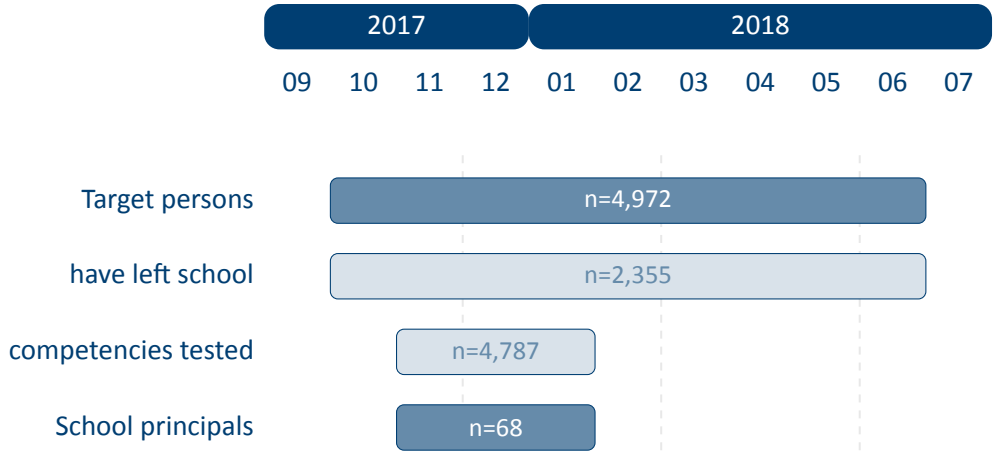

Figure 12: Field times and realized case numbers in wave 9

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Grade 12 students*

    Sample Participants who continue to be willing to participate in the panel

    Modus Written questionnaires and competence tests completed in class context (PAPI)
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted personal interviews (CAPI) with competence testing or Computer-assisted telephone interviews (CATI) without competence testing as alternative plus subsequent Online interviews (CAWI) for selected participants
  - *Individually tracked students*

    Size 844 students in the individual tracking with survey data

    Modus Computer-assisted personal interviews (CAPI) with competence testing or Computer-assisted telephone interviews (CATI) without competence testing as alternative plus subsequent Online interviews (CAWI)

- **Context persons**
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires (PAPI)
- **Data collection**
  - *Commercial survey institutes*

    School context, PAPI IEA DPC–IEA Data Processing and Research Center, Hamburg

    Individual tracking, CAPI/CATI/CAWI infas–Institute for Applied Social Sciences, Bonn

2.4.10 Wave 10: 2018/2019

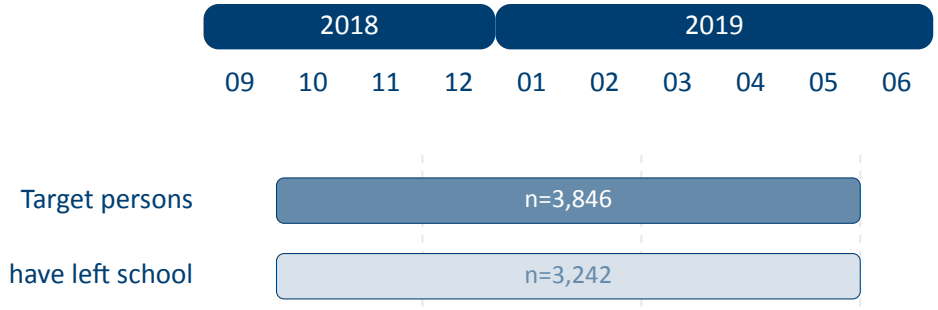

Figure 13: Field times and realized case numbers in wave 10

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted telephone interviews (CATI) or Computer-assisted personal interviews (CAPI) as alternative plus subsequent Online interviews (CAWI) for selected participants, no competence testing

  - *Individually tracked students*

    Size 604 students in the individual tracking with survey data

    Modus Computer-assisted telephone interviews (CATI) or Computer-assisted personal interviews (CAPI) as alternative plus subsequent Online interviews (CAWI), no competence testing

- **Data collection**
  - *Commercial survey institute*

    Individual tracking, CATI/CAPI/CAWI infas–Institute for Applied Social Sciences, Bonn

2.4.11 Wave 11: 2019/2020

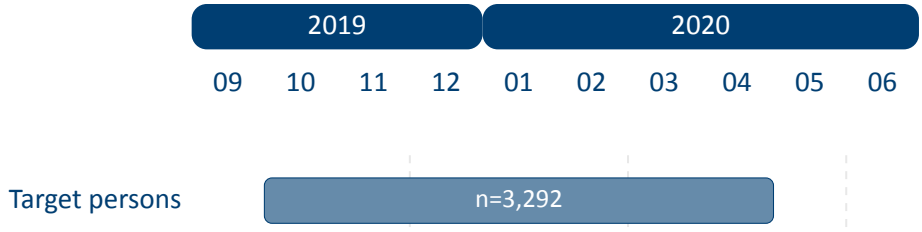

Figure 14: Field times and realized case numbers in wave 11

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted telephone interviews (CATI) or Computer-assisted personal interviews (CAPI) as alternative plus subsequent Online interviews (CAWI) for all participants, no competence testing
- **Data collection**
  - *Commercial survey institute*

    Individual tracking, CATI/CAPI/CAWI infas–Institute for Applied Social Sciences, Bonn

2.4.12 Wave 12: 2020/2021

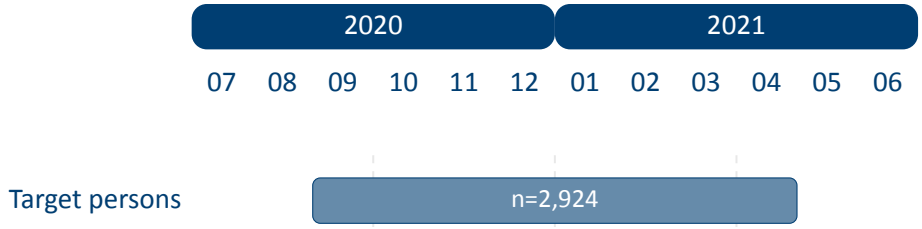

Figure 15: Field times and realized case numbers in wave 12

- **Target persons** 5th graders at panel start 2010/11, 7th graders at refreshment start 2012/13
  - *Individually tracked school-leavers*

    Sample Participants who have left the general education system and continue to be willing to participate in the panel

    Modus Computer-assisted telephone interviews (CATI) plus subsequent Online interviews (CAWI) for selected participants, no competence testing
- **Data collection**
  - *Commercial survey institute*

    Individual tracking, CATI/CAWI infas–Institute for Applied Social Sciences, Bonn

3 General Conventions

The compilation of NEPS Scientific Use Files follows two general paradigms on how to edit the source data (i. e., the data that is delivered to the LIfBi Research Data Center by the survey agencies). There may be exceptions to these principles that are explicitly noted in the respective documentation material.

The first and foremost paradigm in creating NEPS Scientific Use Files is the one of unaltered data. Wherever possible, the data editing procedures do neither change nor destruct the content of the original data. We consider this to be the basis for preserving the full research potential of the collected data. For this reason, no corrections are made during the data editing process to ensure the content validity of the source data. As a consequence, this means that the Scientific Use File may contain implausible values, unless corresponding controls were already provided in the survey instrument. Only in rare cases, in which the responsible developers of a variable require the removal of clearly implausible information, these values are replaced by the special missing code “implausible value removed” (–52, see Table 6). The most prominent (and only systematic) exception to this paradigm concerns the recoding of open responses that could originally have been recorded directly as closed responses (see section 3.4 for details). NEPS Scientific Use Files are equipped with a dataset `EditionBackups` that contains backup information for all content that has been modified by such recoding procedures (see section 4.5.3 for details).

The second paradigm is to integrate the data as much as possible without compromising the usability of the Scientific Use File. The underlying assumption is that for a vast majority of users it is far more comfortable to reduce already integrated data for a specific analysis as opposed to correctly compile the relevant information from scattered source data themselves. In the end, each Scientific Use File contains a few dozen integrated panel and spell datasets according to a general structure (see section 4.3 and section 4.4 for details), even if the compilation is based on several hundred separate source files.

There are additional conventions for the data structure of all NEPS Scientific Use Files. The aim of this overall structuring is to ensure a maximum of consistency between the data of the different starting cohorts. Thus, a researcher who is familiar with the data logic of a particular NEPS cohort should be able to immediately recognize this structure when starting to work with data from another NEPS cohort. The conventions described in the following sections apply equally to Starting Cohort 3, although some of the examples refer to other NEPS starting cohorts.

3.1 File names

The naming of the data files in NEPS Scientific Use Files follows a series of rules that are summarized in Table 3. The different elements are concatenated with an underscore (_) to generate the complete file name.

Table 3: Naming conventions for NEPS file names

Element	Definition
SC[1-6]	**Indicator for the starting cohort** 1 = Newborns 2 = Kindergarten 3 = Fifth-grade students 4 = Ninth-grade students 5 = First-year university students 6 = Adults
[filename]	**Meaning of the file name** *Prefix*: x = cross-sectional file; sp = spell file; p = panel file *Keyword*: indicates the content of the corresponding file (e. g., data file xTarget contains cross-sectional data from the target questionnaire; spSchool contains spell data from the school history) File names of generated datasets do not have a prefix and always start with a capital letter (e. g., CohortProfile, Weights)
[D,R,O]	**Indicator for the confidentiality level** D = Download version R = Remote access version O = On-site access version
[#]-[#]-[#](_beta)	**Indicator for the release version** *First digit*: the main release number is incremented with every further wave in the Scientific Use File; e. g., the first digit 5 implies that data of the first five survey waves are included in the release *Second digit*: the major update number is incremented with every bigger change to the Scientific Use File; major updates affect the data structure, so updating the syntax files may be necessary *Third digit*: the minor update number is incremented with every smaller change to the Scientific Use File; minor updates affect the content of cells, so updating the syntax files is not necessary _beta: this suffix indicates a preliminary Scientific Use File release which allows users to test the data before the main release; the beta release is no longer available after the main release

For instance, the file SC3_CohortProfile_D_12.0.0.dta refers to the *CohortProfile* data of *Starting Cohort 3* in its *Download* version of the Scientific Use File release *12.0.0*.

3.2 Variables

The naming conventions for variables in NEPS Scientific Use Files aim to ensure maximum consistency both between the panel waves and between the starting cohorts. The names also refer to different characteristics and thus allow the data user an orientation regarding the contents of the variables. The principles of these naming conventions are exemplified in Figure 16. It has to be noted that a separate nomenclature is used for variables from competence measurements. Section 3.2.1 offers a detailed description of the general naming conventions for NEPS variables; the logic of naming competence variables is introduced in section 3.2.2.

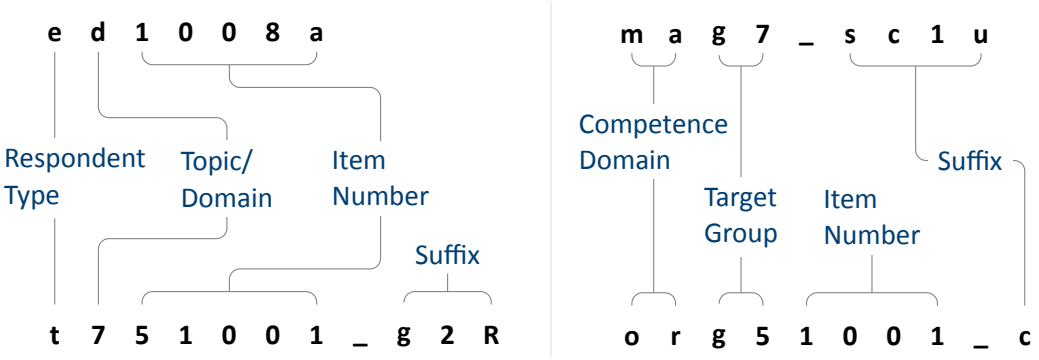

Figure 16: General variable naming (left) and competence variable naming (right)

3.2.1 Conventions for general variable naming

A variable name consists of up to four elements: the respondent type, the domain of information, an item number, and an optional suffix providing further information.

Table 4: Conventions for variable names

Digit	Description
1	**Respondent type** Indicator to which group of respondents the variable refers; note that variables related to the target person start with `t` even if the target person was not the actual informant (e. g., generated variables, list data from schools/kindergartens) `t` = Target person `p` = Parent of target person `e` = Educator/childminder `h` = Head/manager of institution (information about school/kindergarten)

(...)

Table 4: (continued)

Digit	Description
2	**Topic/domain** Indicator to which theoretical dimension or educational stage the variable refers 1 = Competence development 2 = Learning environments 3 = Educational decisions 4 = Migration background 5 = Returns to education 6 = Interest, self-concept and motivation 7 = Socio-demographic information a = Newborns and early childhood education b = From kindergarten to elementary school c = From elementary school to lower secondary school d = From lower to upper secondary school e = From upper secondary school to higher ed./occ. training/labor market f = From vocational training to the labor market g = From higher education to the labor market h = Adult education and lifelong learning m = Corona variables s = Basic program x = Generated variables
3–7	**Item number** Indicator for the item number which typically consists of four numeric characters plus one alphanumeric character
8–11	**Suffixes** (optional, see below) Indicator for several types of variables; separated from the previous characters by an underscore

Suffixes

- *Generated variables:* The _g# suffix indicates a generated variable; the running number after _g is in most cases a simple enumerator (e. g., _g1). Since scale indices are generated by a set of other variables, they are also identified by a _g# suffix. Note that scale indices are named after the first of the set of variables from which they were generated. In this case, numbering is only relevant if the first variable is identical for several scale indices. The number after _g is in most cases a simple enumerator. However, there are two types of generated variables that

assign specific meanings to digits, namely regional and occupational variables. The former are based on the Nomenclature of Territorial Units for Statistics (NUTS):

- g1: Indicator for East or West Germany
- g2: NUTS level 1 (federal state/Bundesland)
- g3: NUTS level 2 (government region/Regierungsbezirk)
- g4: NUTS level 3 (district/Kreis)

Generated variables for occupational classifications and prestige indices (see also section 3.4):

- 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-88 (International Socio-Economic Index of Occupational Status 1988)
- g6: SIOPS-88 (Standard International Occupational Prestige Scale 1988)
- g7: MPS (Magnitude Prestige Scale)
- g8: EGP (Erikson, Goldthorpe, and Portocarero's class categories)
- g9: BLK (Blossfeld's Occupational Classification)
- g14: ISEI-08 (International Socio-Economic Index of Occupational Status 2008)
- g15: CAMSIS (Social Interaction and Stratification Scale)
- g16: SIOPS-08 (Standard International Occupational Prestige Scale 2008)

- *Versions of variables:* If question formulations, interviewer instructions, etc. change between panel waves to such an extent that sufficient meaning equivalence is no longer guaranteed, the answers to these questions are stored in different versions of a variable. The data for the latest and most current version of a question are provided under the variable name without any version suffix. Previous item versions are identified by `var_v1` for the data before the question was modified for the first time, `var_v2` for the data before the question was modified for a second time, and so on.
- *Harmonized variables:* The suffix `var_ha` indicates a harmonized variable in which common information from different versions of a variable is integrated. This is often done by aggregating detailed value characteristics into common superordinate categories. In other words, a harmonized variable reflects the lowest common denominator of information from a variable and its version(s).

- *Wide format variables:* The `_w#` suffix indicates variables that are stored in wide format. Note that this suffix does not necessarily imply a wave logic. The presence of a set of variables `var_w1`, `var_w2`, …, `var_w10` may mean that there are up to 10 values for this variable per person or episode. This is the case, for example, if the corresponding item in the survey instrument was repeatedly measured in a loop. Another example concerns the date of the competence measurement within a survey wave if it took place on two different days.
- *Confidentiality level:* The `_D`, `_R`, or `_O` suffix indicates variables that have been modified during the anonymization process (see section 1.4). The suffix `_O` signalizes that data in this variable is only available via on-site acces; `_R` refers to variables where access to detailed information is only possible via RemoteNEPS and on-site stay; and `_D` means that data in this variable has been extracted from the corresponding `_O` or `_R` variable to make at least some information available in the download version of the Scientific Use File. The confidentiality suffixes stand either alone (e. g., country of birth: `t405010_R`) or in combination with other suffixes (e. g., district of place of birth: `t700101_g3R`).

3.2.2 Conventions for competence variable naming

The naming of variables from competence measurements and direct measures follows an alternative logic. In contrast to other data files, the competence datasets (`xTargetCompetencies` and `xPlausibleValues`, plus `xDirectMeasures` in Starting Cohort 1) are structured in wide format; that is, all values for a single respondent are represented in one row of the data matrix. Thus, the integration of information from several competence domains collected across several survey waves requires specific conventions for variable naming. Competence variables are characterized by three name components and supplementing suffixes. The first component indicates the competence domain of the measurement (two characters, e. g., `vo` for vocabulary). The second part identifies the target group and the survey wave or class level in which the measurement was first used (two or three characters, e. g., `k1` for kindergarten children during the first wave). The target group identification does not necessarily indicate the cohort or testing wave of the measurement. Please refer to the explanations in the next section for the special features of repeatedly used test items. Some competence measurement are not designed for specific age groups, but are implemented unmodified in different cohorts and testing waves. In these cases the target group is defined as `ci` (cohort invariant). The third component denotes the item number. Table 5 contains all specifications of a competence variable name.[7]

The additional suffixes inform about the mode of test execution if more than one survey modus has been applied for a measurement and about the sort of item score and overall competence score. There is a distinction between scored items named `[varname]_c` and scored partial credit-items named `[varname]s_c`. The latter is relevant if more than one correct solution is possible (e. g., value 0 = “0 out of two points”, value 1 = “1 out of two points”, value 2 = “2 out of two points”), whereas the former is applied for dichotomous solutions (value 0 = “not solved”,

7 The variables generated from the competence data in the additional dataset `xPlausibleValues` follow the same naming logic – with a uniform suffix `_pv#` after the first two parts of the naming convention.

value 1 = “solved”). In addition to the item scores, several aggregated scores are provided for competence measurements. They are indicated by `_sc[number]` and a few special suffixes for Starting Cohort 1. A letter appended to the suffix indicates that more than one aggregated score for a competence measurement is available (e. g., `_sc3a`, `_sc3b` for different sum scores of any test). Detailed descriptions on how the aggregated competence scores were estimated can be found in the domain-specific documentation reports. The last part of Table 5 shows all possible suffixes and their meanings.

Table 5: Conventions for competence variable names

Part I: Competence Domain (2 chars)

ba	Business administration and economics
bd	Backwards digit span: Phonological working memory
ca	Categorization: SON-R subtest
cd	Cognitive development: Sensorimotor development
de	Delayed gratification: Executive control
dg	Domain-general cognitive functions (DGCF): Cognitive basic skills
ds	Digit span: Phonological working memory
ec	Flanker task: Executive control
ef	English foreign language: English reading competence
fa	FAIR: Concentration abilities
gr	Grammar: Listening comprehension at sentence level
hd	Habituation-dishabituation paradigm
ic	Information and communication technology literacy (ICT)
ih	Interaction at home: Parent-child interaction
ip	Identification of phonemes: Phonological awareness
li	Listening: Listening comprehension at text/ciscourse level
lk	Early knowledge of letters
ma	Mathematical competence
md	Declarative metacognition
mp	Procedural metacognition
nr/nt	Native language Russian/Turkish: Listening comprehension
on	Blending of onset and rimes: Phonological awareness
or	Orthography
re	Reading competence
ri	Rimes: Phonological awareness
rs	Reading speed
rx	Early reading competence
sc	Scientific competence
st	Scientific thinking: Science propaedeutics
vo	Vocabulary: Listening comprehension at word level

(...)

Table 5: (continued)

Part II: Target Group (1 char)**, followed by wave or grade** (1-2 digits)

n#	Newborns in wave #
k#	Kindergarten children in wave #
g#	Students at school in grade #
s#	University students in wave #
a#	Adults in wave #
ci	Cohort invariant (for instruments administered unchanged in all cohorts)

Part III: Item number (3-4 chars)

For some competence domains, these item numbers follow a certain scheme, but for most competence domains they only indicate the different items

Part IV: Suffixes (starting with an underscore)

_pb	Paper-based test modus (proctored)
_cb	Computer-based test modus (proctored)
_wb	Web/Internet-based test modus (unproctored)
_c	Scored item variable (s_c for partial credit-items)
_sc1	Weighted likelihood estimate (WLE) [a] [b]
_sc2	Standard error for the WLE [b]
_sc3	Sum score
_sc4	Mean score
_sc5	Difference score (for procedural metacognition)
_sc6	Proportion correct score (for procedural metacognition)
_p	Maximum value for an item (only in Starting Cohort 1)
_b	Minimum value for an item (only in Starting Cohort 1)
_m	Mean value for an item (only in Starting Cohort 1)
_s	Sum value for an item (only in Starting Cohort 1)
_n	Number value for an item (only in Starting Cohort 1)

[a] WLEs and their standard errors are estimated in tests that are scaled based on models of Item Response Theory (cf. Pohl and Carstensen, 2012).

[b] WLEs and their standard errors are corrected for test position; uncorrected WLEs and standard errors are indicated by an additional u in the suffix (_sc1u, _sc2u).

Identification of repeated test items

In some competence measurements identical items are implemented in different testing waves (e. g., mathematics). Identifying repeatedly measured test items in NEPS data can be easily done by looking for competence variables with an identical word stem. If the same test item is surveyed in different survey waves or starting cohorts, the variable name is equiped with an

additional suffix. It is important to know that the two or three characters for the target group (second part of the variable name) always indicate the wave or cohort in which the item was initially used. The word stem is then fixed and does not change when the item is used again in later waves or other cohorts. If the variable name does not contain a suffix for repeated use, then the second part of the word stem refers to the target group of the realized measurement. However, if the variable name includes a suffix for repeated use, then the values of the variable do not refer to the target group according to the word stem, but to the target group according to the suffix. The suffix that points to the repeated use consists of two parts: The first element indicates the starting cohort of current item administration and the second element indicates the cohort or testing wave of current item administration.

The following example illustrates this logic: The competence variable `vok10067_sc2g1_c` is a vocabulary item (`vo`) that was initially measured during the first kindergarten survey wave (`k1`). However, the values in this variable reflect the scored measurements of this item´s repeated use among the target persons of Starting Cohort 2 in the course of the survey wave in grade 1 (`_sc2g1`), and thus two years after the first measurement.

3.2.3 Labels

As a rule, the seven-digit variable names are not sufficient to uniquely identify the respective contents of the variables and to differentiate sufficiently between items. All variables therefore have *variable labels* for more detailed description. In addition, most variables contain *value labels* for the respective value characteristics. All information is available in German and English and is typically displayed directly in the editor of the statistics program, e.g. for frequency calculation or when searching the data (applies to SPSS and Stata, see also section 1.3). For users of R, see section B.1 for hints on this.

In addition to the variable and value labels, the datasets also contain extended characteristics for variables. These include the question text from the survey instrument, any associated interviewer instructions and filter conditions, as well as other meta information. All extended features can be accessed directly within data files. Stata users apply the `infoquery` command for this, which is part of the *NEPStools* package (see section 1.8). SPSS users will find the additional meta information in the “Variable View” at the end of each variable line.

As explained in more detail in section 4, NEPS data from different waves are integrated as much as possible. For panel data, this primarily means that many variables contain information from multiple waves. In most cases of such a data integration, the meta information between the waves does not change. However, if there are changes to the meta information of a repeatedly measured item, and if these changes are not significant enough to store the information in separate variables, the assignment of meta information follows a general rule: **All meta information available in a dataset always corresponds to the most recently instrument in which the respective item was used.**

A concrete example is the adaptation of interviewer instructions or question texts from the informal salutation ("Du") to the formal salutation ("Sie"). Since these changes are not expected to have any effect on how a question is answered, the corresponding values across multiple waves get integrated into one variable. If you request the meta information of such a variable in the dataset, the wording of the latest item formulation will be displayed (in the given example with the formal salutation "Sie"). In case of uncertainties regarding the continuity of meta information of a variable across different waves, we recommend to consult the respective *survey instruments* for the individual waves.

3.3 Missing values

The NEPS data contain various missing codes to differentiate between various types of missing values. All missing codes have negative values or are defined as system missing. Depending on the statistics program used, you must ensure that these codes are processed correctly. In the offered SPSS datasets, the missing codes are already defined as missing values. When using Stata, the missing codes must first be excluded from the analyses by the user as missing values. For this purpose the command `nepsmiss` is available in the *NEPStools* package (see section 1.8). The general recommendation is to always carefully check the frequency distributions of the relevant variables before running an analysis. The three main types of missing codes are summarized in table 6 and described below.

Table 6: Overview of missing codes

Code	Meaning	Note
Item nonresponse		
–94	not reached	only relevant for instruments with time restrictions (e. g., competency test measures)
–95	implausible value	assigned by the survey agency (e. g., multiple answers to a one-answer question in PAPI mode)
–97	refused	as default answer option to the question
–98	don't know	as default answer option to the question
–20,…,–29	*various*	item-specific missing with informative value label (e. g., "no grade received" for question about school grades)
Not applicable		
–54	missing by design	question not included in (sub)sample-specific instrument (e. g., not asked in all waves)

(…)

Table 6: (continued)

Code	Meaning	Note
–90	unspecific missing	in PAPI mode (e. g., question not answered, empty field)
–91	survey aborted	respondent quit interview, in CAWI mode
–92	question erroneously not asked	question not asked by mistake, in CAWI and CATI
–93	does not apply	as default answer option to the question
–99	filtered	filtered out question, in other than CATI/CAPI mode
.	*system*	filtered out question, in CATI/CAPI mode
Edition missings (recoded into missing)		
–52	implausible value removed	only at the request of the responsible item developers
–53	anonymized	sensitive information removed (e. g., country of birth of parents in the download version)
–55	not determinable	not sufficient information to generate the variable value (e. g., net household income `t510010_g1`)
–56	not participated	in case of unit nonresponse, only used in certain datasets

Item nonresponse: The first type of missing codes occurs when a person has not (validly) replied to a question.

- The most common cases of item nonresponse are “refused” (–97) answers and “don’t know” (–98) answers.
- Missing values specified by the survey agency due to an incorrect use of the instrument are coded as “implausible value” (–95).
- Within the competence data, there is a special missing code indicating that a question or test item was “not reached” (–94) due to time constraints or other test setting restrictions. It usually signals that the respondent had to quit the test somewhere before this point.
- Other missing codes refer to various categories of “item-specific nonresponse” (–20, …,–29) such as –20 for “stateless” in the citizenship variable `p407050_D`.

Not applicable: The second type of missing codes occurs when an item does not apply to a respondent.

- The code "missing by design" (–54) is assigned when respondents in a (sub)sample have not been asked the respective questions. This is usually the case if the administered survey instrument contains (sub)sample-specific questionnaire modules. The code is also used for the more general case where values of a variable are not available due to the design of the survey (e. g., measurement rotation with either easier or heavier test tasks).
- If the respondent him-/herself or the interviewer indicates that a particular question is not applicable to the person, the missing value is coded as "does not apply" (–93). If, on the other hand, filtering takes places automatically via the survey instrument, the coding of the filtered out questions depends on the survey mode: in CATI and CAPI interviews, a system missing value (`.`) is assigned for this; in all other modes the respective code is "filtered" (–99).
- Missing values that cannot be assigned to any of the above categories are coded as "unspecific missing" (–90). This missing code usually occurs in PAPI questionnaires when a respondent has not answered a question for unknown reasons.

Edition missings: The third type of missing codes is defined in the process of data preparation for the Scientific Use File.

- If in the data edition process certain values which are not considered to be meaningful are requested to be removed, the missing code "implausible value removed" (–52) is assigned in their place. As a rule, however, all values from the field instruments are included in the Scientific Use File without further plausibility checks (see section 3). Only in exceptional cases, when the responsible item developers explicitly recommend a removal of implausible answers, this missing coding is done.
- Sensitive information that is only available via remote and/or on-site access is encoded in the more anonymized data access option as "anonymized" (–53).
- In general, coding schemes are used to generate variables (e. g., occupational coding; see section 3.4). However, if the information from the original data is not sufficient to generate a suitable value, the missing code "not determinable" (–55) is used instead.
- If a person was not present during the interview or did not complete a questionnaire at all, even though it was administered to the person, the concerning variables receive the code "not participated" (–56). This missing code is special in the sense that target persons for whom no survey data at all are available for a certain wave (e. g., due to illness) are usually not included in the corresponding datasets. This missing code is only used in the special cases of datasets that integrate several waves in wide format (e. g., `xTargetCompetencies`) or that also contain observations for non-participating persons in a wave (e. g., `CohortProfile`).

3.4 Generated variables

Coding and recoding of open responses

At various points in the NEPS survey instruments there are so-called open questions where respondents can or should enter their answers as text. A typical example is information about occupation.

The open text format allows respondents to specify anything they want. A practical way to deal with the resulting string information is to code and recode the information for further processing and later analyses. In general, coding describes the process of assigning one or more codes from selected category schemes to the string information, e. g. the classification of occupational data according to DKZ (database of documentation codes, *Datenbank der Dokumentationskennziffern*) or WZ (classification of economy branches, *Klassifikation der Wirtschaftszweige*).

The term "recoding" is used here to describe the process of assigning a code from an already presented closed answer scheme. This usually applies to semi-open question formats where respondents enter a text under the category "other", but which can be assigned ad hoc to one of the given closed answer categories. The recoding therefore does not define any new codes; the presented answer scheme of the respective question is not extended.

The most common and comprehensive coding scenarios in the fields of occupation, education, branches, courses, and regional information are processed by the Research Data Center (FDZ-LIfBi) itself. Other coding tasks are distributed among the responsible departments at the LIfBi in Bamberg and the partners in the NEPS consortium.

Derived scales and classifications

The (re-)coding of open answers or string entries into primary classifications (such as DKZ2010 or WZ08) is a first and essential step towards making this information available within the NEPS Scientific Use Files in a user-friendly and analyzable way. The standardized derivation of further classifications or scales, especially in the area of educational qualifications and occupational titles, is a second and no less important step. At least three types and objectives of derivations can be distinguished:

- Derivations from primary classifications (and originated from string entries/open answers) into other classifications that function as a standard schema in other studies or international comparisons, e. g. ISCO instead of KldB in the field of occupations
- Derivations from primarily closed response schemes into general classifications and schemes using auxiliary information, e. g. ISCED or CASMIN from school certificate and training data plus additional information on the type of school/training
- Combination of the two types, e. g. EGP class scheme via derived ISCO classification plus information on self-employment and supervisory status

Figure 17 shows the derivation paths for several occupational scales and schemes provided in the NEPS. A detailed description of the standard derivations for educational attainment (ISCED, CASMIN and Years of Education) can be found in the corresponding documetation report by Zielonka and Pelz, 2015.

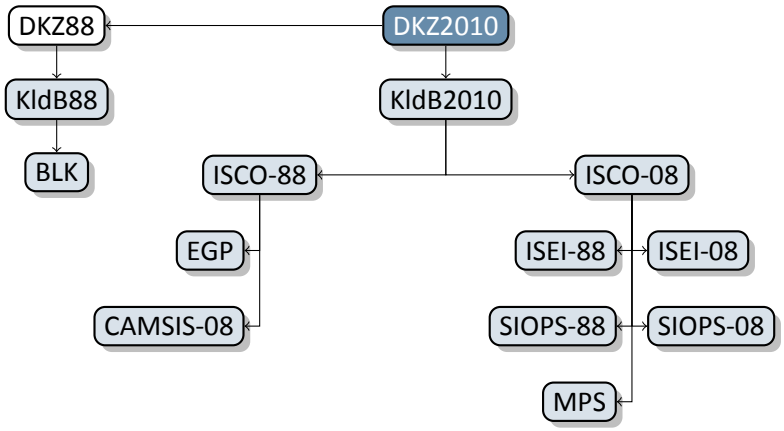

Figure 17: Derivation paths for several occupational scales and schemes provided in the NEPS

4 Data Structure

4.1 Overview

The longitudinal NEPS study is a complex research database. It is the result of extensive data edition processes with the aim of organizing the information in a well-structured, reproducible and user-friendly way, while at the same time preserving a maximum level of detail in the data. To facilitate the handling of the data, a number of additionally generated variables and datasets is included in the Scientific Use File.

In principle, all information collected in the course of a panel wave is appended to the information from previous waves in the corresponding data file, together with the required identifiers. Data files containing panel information from several waves are denoted with a *p* at the beginning of the file name. For example, the `pTarget` file contains information from the target persons' interviews with one row in the dataset representing the information of one individual in one wave (see section 4.3).

This convention, however, does not apply to all longitudinal information in the NEPS. For example, there are competence measurements that were repeatedly carried out with the same target persons. However, since the content of competence tests varies over time, the corresponding data is structured in *wide format* (see section 3.2.2). Such cross-sectionally structured data files with one row representing information of one individual from all waves are marked with an *x*.

Another type of longitudinal data structuring refers to episode or spell data (see section 4.4). For the information collected prospectively and retrospectively by using iterative question sets, the Scientific Use File provides numerous life area-specific spell datasets (see also section 5). These datasets are marked by a preceding *sp*. An example is the file `spEmp`, which informs about current and former episodes of employment.

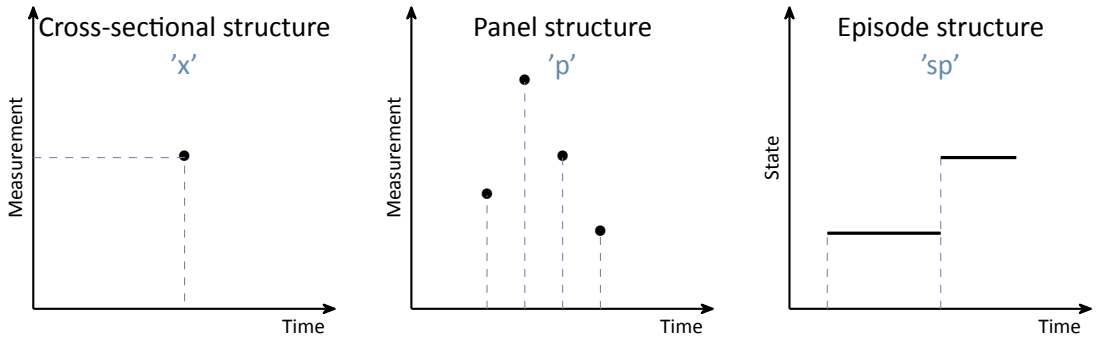

Figure 18: Different types of data structures

In addition to the interview, competence and episode data surveyed from the respondents, there are so-called paradata and derived information available. The respective data files can be identified by the leading capital letter in the name (e. g., `Weights`, `TargetMethods`, `Biography` or `CohortProfile`, see figure 20).

4.2 Identifiers

The multi-level and multi-informant design of the NEPS together with the provision of information in different files requires the use of multiple identifiers. The following identifier variables are relevant in Starting Cohort 3 for merging data from different datasets:

ID_t identifies a target person. The variable `ID_t` is unique across waves and samples; it is also used uniquely in each Starting Cohort.

wave indicates the survey wave in which the data was collected.

ID_i identifies the respective educational institutions such as kindergartens or day care centers, schools, universities, etc. The variable `ID_i` is unique across waves and starting cohorts.

splink uniquely identifies episodes/spells across all datasets within each person. It is used to link data from `Biography` with `Education` or episode modules such as `spVocTrain`.

ID_cc, ID_cg, ID_cm identifies the class, the German course, and the mathematics course within a wave. These identifier variables are **not** unique across waves. Regarding the assignment of students, it should be noted that the primary organization is the class and not the course context. This means that a large proportion of the three identifier variables have the same value for a given target person within a given wave. However, there are also deviations from this pattern.

ID_e uniquely identifies an educator/teacher across waves. This identifier variable can be used to merge data from educators/teachers with observations from children/students. However, it is not possible to merge the data directly with the `ID_t` (e. g., in the `CohortProfile` dataset). The linking with data of the target persons or parents or institutions requires the "path" via the class or course identifier (`ID_cc` in `pCourseClass` or `ID_cg` in `pCourseGerman` or `ID_cm` in `pCourseMath`). A concrete example of the procedure is given in section 4.5.9.

There are further identifier variables to indicate a target person's membership in a particular test group (`ID_tg` in `CohortProfile`, not applicable to all starting cohorts) or to indicate the interviewer who conducted the respective interview (`ID_int` in `Methods` datasets). These identifiers are less relevant for the merging of information from different datasets and negligible for most empirical applications.

4.3 Panel data

In general, all information from the latest survey wave is appended to the already existing information from previous waves (as far as possible). This kind of data preparation generates integrated panel data files in a *long format* as opposed to providing one separate file per wave (where each file contains only the information from a single wave). When working with the integrated NEPS panel data, the following points are important to be considered:

- A row in the dataset contains the information of one respondent from one survey wave.
- More than one variable is needed to identify a single row for uniquely selecting and merging information from different datasets. Usually, `ID_t` and `wave` are the relevant identifiers.
- Although not all questions were administered in each survey wave, the data structure contains cells for all variables and waves. If no data is available, e. g., because a question was not asked in a wave, the corresponding cells are filled with a missing code (see section 3.3).
- If information about a variable has been repeatedly surveyed from one individual across multiple waves, the corresponding data is stored in multiple rows in the dataset.

The long format is usually the preferred data structure for the analysis of panel information. However, cross-sectional information is often required as well in analyses, e. g., because it depicts time-invariant characteristics or was collected only once for other reasons. In most scenarios, the relevant set of variables might not have been measured in a single wave. Therefore, the data cannot be analyzed together straightaway because it is stored in *different rows* of the dataset. Cross-tabulating these variables in their current state results in an L-shaped table in which all observations of one variable fall into the missing category of the other variable and vice versa. The best way to deal with this issue depends very much on the intended analysis and the methods used. The two typical procedures are:

- The integrated panel data file is split into wave-specific subfiles so that each dataset contains only information from one wave. The relevant information from these subfiles is then merged together by using only the respondent's identifier (`ID_t`) as key variable. The `wave` variable is not needed here and remains neglected. Before this step, variables may need to be renamed to make them wave-specifically identifiable. The result is a dataset with a cross-sectional structure in which the information of one respondent is summarized in one single row (wide format). Stata's *reshape* command (and similar tools in other software packages) basically follow this strategy.
- Alternatively, the panel structure is retained and the values from observed cells of a variable are copied into the unobserved cells of this variable. For example, if the place of birth was only surveyed in the first wave, the corresponding value can be copied into the respective cells of the respondent's other waves. This method is particularly useful for time-invariant variables (e. g., country of birth, language of origin), that are usually collected only once in a panel study.

4.4 Episode or spell data

Episode or spell data are particularly challenging to handle. The following explanations help to understand this data format and to deal with it in a meaningful and appropriate way. For further details please refer to the "Special Issues" in section 5.

In episode data, there is one row for each episode that was captured during the interview. Usually, a start and an end date describe the duration of the episode. The remaining variables in spell datasets provide additional information about that episode. These descriptors are related to the particular episode and fill it with content, so to speak. It means (especially for time-variant variables like education or occupation or employment) that the respective values indicate the status *at the time of the episode*, which is not necessarily the current status valid nowadays (or at the time of the interview). To give an example, in the dataset `spEmp` there is a period of time for a particular respondent during which she or he worked in a particular job without interruption. If this person changed to a new job, this defines a new episode stored in a new data row. Further changes in this context may also lead to new episodes, e. g., a change of the employer or the conclusion of a new employment contract – but not if the salary, working hours or other characteristics (possible descriptors) of the respective job change. Episodes can be understood as the smallest possible units of one's life history, in this case the employment biography. Several relevant changes in such a biographical area are reflected in several new data rows.

To make this clear: The number of episodes is per se independent of the survey wave. During an interview (one wave) there might be a number of episodes recorded (several rows) or no episode at all (no row). The dates given for an episode relate to that episode, whereas the wave indicator relates to the interview date. The two can overlap, but do not have to. Data users should consider both entities – `spell` and `wave` – to be independent of each other. In exceptional cases, it might be important to know when the information about an episode was collected. Beyond that, however, the variable `wave` can be ignored in the episode data. In particular, the `wave` variable should **not** be used to merge episode data with panel data in the long format. Since episode data may contain multiple (or no) rows per survey wave and target ID, and panel data contain exactly one row for each survey wave and target ID, such a merge will result in converting the panel data to an episode structure. The result of this kind of transformation is no longer analyzable in a meaningful way. A better approach is to aggregate the episode data to one piece of information either for each interview date (e.g., number of jobs since the last interview) or for the entire life course (e.g., highest educational attainment), so that only one row per survey wave and respondent is left for the merging process.

In addition to (time-dependent) episode data such as jobs, which we call *duration spells*, there are two other types of episode spells in the NEPS data:

- Occurring events or the transition from one state to another (e. g., change of marital status, change of educational level) are recorded in *event spells* with one row describing one state.
- The existence of children, partners, etc., is recorded in *entity spells* with one row per entity.

Regardless of the type of episode, at least two variables are necessary to identify a single row in the data file, namely the respondents' identifier `ID_t` and an numerator for the episode, event or entity such as `spell` or `child`. More detailed information on the available identifier variables can be found on the respective data file descriptions in section 4.5.

4.4.1 Edition of the life course

The life course data in all NEPS starting cohorts mainly consists of information on episodes of school attendance, participation in vocational preparation measures and vocational training, university education, as well as of compulsory or voluntary services, employment and unemployment, and parental leave. We refer to these activities as *main activities*. The episodes are grouped by type and recorded in separate modules (see also section 5.1 for details). The aim of this recording is to capture chronologically complete life histories across key biographical areas of the respondents. This goal is supported by two data-guided measures:

Data edition during the interview

The first step of editing the life history information takes place during the interview. The episodes reported by the respondent are summarized by the instrument and put into a chronological order. They are then checked for gaps and overlaps. Their clarification is made cooperatively by the interviewee and the interviewer with the help of the so-called *check module* (Hess et al., 2012).

If chronological *gaps* are identified, they are subsequently closed by recording additional episodes with regard to the above-mentioned main activities. If there is no suitable main activity for a gap, the respondent can close it with a "gap activity". Moreover, gaps can be filled by adjusting the start and end dates of the episodes between which the gap exists (see also section 5.3.9).

Chronological *overlaps* of episodes are also reviewed together with the respondent. This may lead to an adjustment of the dates of the episodes involved in the overlap. For imprecise or missing date information, estimates are calculated where there is reasonable evidence. For example, the rather vague specification "summer" for the starting month of an episode is replaced by the value 7 for "July" and stored accordingly. This allows episodes with incomplete dates to be included in the plausibility test during the interview and to be checked in the overall context of the reported life history (Ruland et al., 2016; Matthes et al. 2005, 2007).

Data edition after the interview

Despite extensive review during the interview with largely complete and chronologically consistent life histories as a result, there might still be minor inaccuracies at the end. For example, one-month overlaps of episodes are not displayed or processed in the check module. The same applies to gaps of up to two months between consecutive episodes. Also, the review can be interrupted or skipped at the request of the respondent. Therefore, a second step of automated editing of biography information takes place after the end of the interview (Künster

2015a, 2015b). The results of this successive data edition concern only the `Biography` dataset. In the spell datasets for the different life domains (e. g., `spEmp`), the information provided by respondents during the interview with regard to the start and end dates of episodes remains unchanged.

- First, one-month overlaps of episodes are removed. Such an overlap occurs when the end date of a previous episode is identical to the start date of the following episode, i.e. the same month was mentioned. In this case, the end date of the previous episode is shortened by one month. The condition for this is that the previous episode is longer than one month. If this condition is not met, the start date of the following episode is shortened by one month. If both episodes have a duration of only one month, the dates remain unedited.
- Second, one- and two-month gaps between consecutive episodes are automatically closed. For a one-month gap, the end date of the previous episode is extended by one month. For a two-month gap, the start date of the following episode is additionally moved forward by one month.
- Finally, chronological gaps in the life history that are larger than two months are closed by inserting new episodes into the `Biography` file. These artificial episodes, labeled as "data edition gap" in the variable `sptype`, completely close larger gaps.

4.4.2 Revoked episodes

To make it easier for respondents to answer the life history modules and to minimize recall errors, information on episodes from previous interviews is preloaded. This information can be subsequently revoked during the current interview. The spell datasets also contain these revocations or contradictions (variables `disagint`, `disagwave`). The reasons for that are manifold; they primarily depend on the information presented to the interviewed person to remember an episode (the exact wording of the episode data collection can be seen in the questionnaires).

Subsequently revoked episodes are marked accordingly in the respective dataset. The information collected again in the current interview is additionally stored as a new episode in the corresponding (more recent) survey wave. That updated episode is **not** marked as a corrected spell. The identification of related spells – original information plus its correction in the subsequent survey wave – is up to the data user. It should be noted that virtually all corrected episodes are *left-censored*. This is because it is technically not possible to specify a start date for an episode in the interview that precedes the last interview. The earliest start date is for episodes that began on the interview date of the last survey.

In addition, there is also the possibility of revoking a reported episode still during the interview. The check module (see section 4.4.1) is also used for this purpose after all current biographical information has been recorded. It ensures that the life course is captured as completely and consistently as possible. As part of the plausibility review within the interview, there is the

option for respondents of correcting and also revoking previously reported episodes. The identification of episodes that were revoked in the check module is possible by the variable `spms` "check module: type of event" and the value -20 "episode revoked in check module"). The addition of new episodes in the check module is indicated in the "episode mode" variable such as `ts23550=4` in `spEmp`).

4.4.3 Subspells and harmonization of episodes

When working with NEPS spell data, there is an important circumstance to consider: Biographical episode data are collected retrospectively. During an interview, respondents are asked about all episodes that have occurred since the last interview (or the first interview, since birth or a certain age). If an episode ended before the time of the current interview, the respondent provides an end date and the spell is completed. Challenges occur when the episode has not ended at the time of the interview, i.e., it is still ongoing.

Such an episode appears in the dataset as *right-censored*. In the next interview, this episode is then preloaded in the course of the "dependent interview" in a way that the respondent can report whether it has been finished in the meantime or whether it still continues. Technically, this results in multiple rows in the data structure, which can be distinguished by the variable `subspell`:

- first data row with initial information about an episode (right-censored) reported in survey wave x (`subspell=0` if this is the only subspell for that episode, `subspell=1` if there are other subspells from later waves)
- second and further data rows for the continued episode, reported in subsequent survey waves x+ (`subspell=2`, `subspell=3`, etc.)

To make it easier for data users to work with these spread episode data, they are also summarized in a data line (record) according to defined rules. This data line reflects the most current information on the episode. This means that for completed episodes, the information valid at the end of the episode is selected and for episodes that were not yet completed at the last interview time, the information valid at the last interview time is selected. We call this process of summarizing information about an episode from different survey waves *episode harmonization*. It is described in detail below.

An episode is defined by the assignment to a respondent (`ID_t`), by the type (e. g., training episode), by the episode identificator (`splink`, which typically consecutively numbers episodes of the same type for a case), and by the start and end date.

If an episode starts and ends within the retrospectively queried time period of a survey wave (spell 1 in interview A, see figure 19), it can be assumed that this episode has been recorded completely with all information. In the corresponding spell dataset of the Scientific Use File, this episode appears in a single data row.

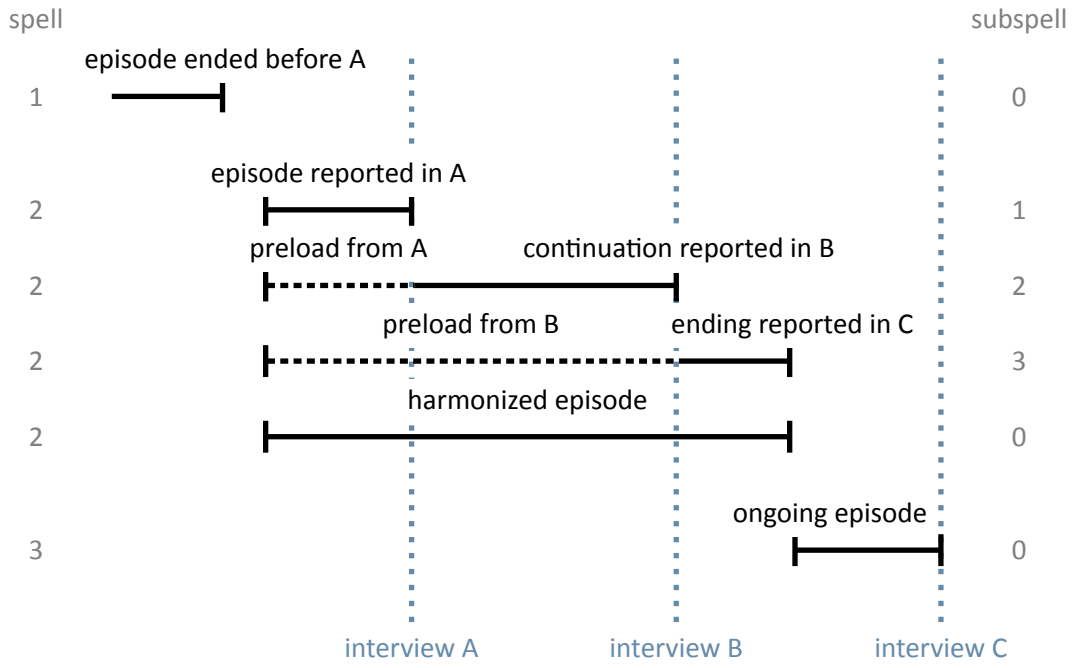

Figure 19: Logic of subspells

However, there are episodes that have not yet finished at the time of the interview, but continue beyond that point. Such episodes are updated in the subsequent survey wave in which the respondent participates. That is, further information about the episode is collected in one or more subsequent waves until the episode is reported as finished (spell 2 in interview B and interview C, see figure 19). In such cases, information about an episode is stored separately in one data row for each survey wave. Accordingly, the information is spread over several data rows and a single data row contains only a subset of information for that episode. The respondent ID is identical in each data row for this episode, as well as the episode ID. The distinction is made by the variable `subspell`, in which the data rows belonging to an episode that was recorded over several survey waves are consecutively numbered (starting with the value 1).

Analogous to episodes that began and ended within the time period of a survey wave (spell 1), the variable `subspell` has a value of 0 also for episodes that were recorded for the first time in the current survey wave and were still ongoing at the day of the interview (spell 3 in interview C, see figure 19).

The sample episodes from figure 19 correspond to the data structure presented in table 7 *before* any episode harmonization.[8] There is only one data row for the first episode. It was completed before the data collection of wave 2, i.e. the information is completely recorded. The value of the variable `subspell` is 0. The second episode is spread over three data rows with information

8 For the sake of convenience, the table only includes data from three consecutive survey waves, conducted in December 2009 (wave=2), 2010 (wave=3), and 2011 (wave=4).

asked in the surveys waves 2 to 4. The values of the variable `subspell` are 1 to 3 according to the consecutive numbering of the sub-episodes. The third episode was recorded in the fourth survey wave. This episode continues, but since only part of the episode has been reported so far, `subspell` is also given the value 0. This value changes as soon as further information about this episode is added in a subsequent survey wave.

Table 7: Data lines of the example case in the SUF before spell harmonization

ID_t	splink	wave	subspell	start_m	start_y	end_m	end_y	ongoing	var1	var2
1	300001	2	0	may	2005	april	2009	no	3	5
1	300002	2	1	june	2009	december	2009	yes	1	.
1	300002	3	2	june	2009	december	2010	yes	.	.
1	300002	4	3	june	2009	july	2011	no	.	8
1	300003	4	0	august	2011	december	2011	yes	2	4

For episodes that span over several survey waves, the same information is not collected in each survey wave. In the wave in which an episode is recorded for the first time, all unchanging core information about it is captured. In the example of training episodes, this includes the start date, the type of training (e. g., vocational training or study), the exact name of the training occupation and some other parameters that distinguish this training from others. In later survey waves, this information is no longer requested when updating this episode. Instead, additional characteristics, such as current pay, are recorded. Once the respondent indicates that the episode has been finished, information about the end is recorded. This is, for example, the achieved completion of a training and, of course, the end date of the episode. Thus, the information about an episode that lasts over several survey waves is divided among sub-episodes (subspells). The number of sub-episodes varies depending on the total duration of the episode or the number of interviews in the course of this duration.[9] To ease the work with updated episodes, the information from the sub-spells of an episode is summarized in an additional data row. Besides the data rows for the sub-episodes, there is one data row that gives an overall view of the entire episode (up to the last interview). This data row represents the *harmonized episode*. Episode harmonization is only used if several subspells from different survey waves are available for the same episode.

Table 8: Data lines of the example case in the SUF after spell harmonization

ID_t	splink	wave	subspell	start_m	start_y	end_m	end_y	ongoing	var1	var2
1	300001	2	0	may	2005	april	2009	no	3	5
1	300002	2	1	june	2009	december	2009	yes	1	.
1	300002	3	2	june	2009	december	2010	yes	.	.
1	300002	4	3	june	2009	july	2011	no	.	8
1	300002	4	0	june	2009	july	2011	no	1	8
1	300003	4	0	august	2011	december	2011	yes	2	4

The data row for the harmonized episode is simply added to the existing data rows for an episode. It is always identified by the value 0 in the variable `subspell`. In the example case, the

9 An update of episodes is only carried out in the Starting Cohort 3 for the following SUF files: `spChild`, `spChildCohab`, `spEmp`, `spGap`, `spMilitary`, `spParentSchool`, `spParLeave`, `spSchool`, `spUnemp`, `spVocPrep`, `spVocTrain`.

additional data row concerns the second episode (`splink=30002`) as a summary of the three sub-episodes (see the highlighted row in table 8). The other two episodes do not have multiple subspells across different survey waves, so harmonization is not necessary or possible.

Since the harmonized spell is a summary of all subspells of an episode, exactly one piece of information must be selected from these subspells for each variable to be transferred to the harmonized spell. In most cases, the rule for selecting the relevant information is obvious. If this is not the case, the following rules are applied:

first For all variables that are filled only at the start of a new episode, i.e. when the episode is first reported, the information from the first sub-episode goes into the harmonized spell, since it can be found only there and is valid for the entire duration of the episode (see `var1` in table 8).

last For information that is newly collected in each survey wave or that is only present in the last subspell of the episode, the information for the harmonized spell is taken from the last subspell (see `var2` in table 8).[10]

first nonmissing The harmonization of most variables follows either the *first* or *last* selection rule. However, there are exceptions. One such exception is when a new question is introduced in the collection of episodes whose variable basically follows the *first* rule, but which is collected in the current survey wave for an episode that is already continuing. In such cases, the information is included in the data for an updated episode, however, not in the first subspell, but in a later subspell. In these cases, the first valid value found in any subspell of an episode is selected.

last nonmissing A similar exception applies to variables that measure a changing state until a defined target state is reached. In the case of employment episodes, for example, this might be the change from a temporary position in a particular job to a permanent position. In cases where a position is temporary at the time of the first recording, the question about the temporary nature of that position is asked each time in subsequent survey waves. This continues until the employment either ends or the status changes to "permanent". Once this change has occurred, the question about a fixed term is no longer asked when the episode is updated later on.[11] Thus, the information about the fixed term of the episode is not necessarily in the first or in the last subspell. Here, the last valid value of a subspell of the episode is relevant. For this reason, the rule *last nonmissing* (last valid value found in the subspells of an episode) is used for harmonization.

10 There is one exception to the use of this harmonization rule: If a question in a life course module is generally not asked in a certain survey wave, then the value of the corresponding variable is set to -54 "missing by design". If the harmonization rule *last* applies to this variable, then the value -54 is **not** transferred to the harmonized episode. Instead, a value different from -54 is searched for in the existing sub-episodes of the episode in question. This value is then stored in the harmonized episode. The same is true for the value -55 "not applicable". The idea is that the value determined in this way is a good estimate of the missing last information for that element of the episode.
11 A reverse change from permanent to temporary within the same job is not considered very realistic.

The Research Data Center at LIfBi protocols which harmonization rule was applied to which variable of life history episodes that have been updated over several survey waves. These harmonization tables are not publicly available, but can be viewed upon specific request.

There is another special aspect regarding the harmonization of episodes: Respondents have the possibility to contradict the update of an episode in the current survey wave in the course of the review of the data in the check module (see section 4.4.1 and Ruland et al., 2016). Only episode types included in this check during the interview are affected (from `spSchool`, `spVocPrep`, `spVocTrain`, `spMilitary`, `spEmp`, `spUnemp`, `spParLeave`, `spGap`). In the case of such a contradiction, the data edition assumes that the subspells recorded in previous waves of the survey contain correct information about this episode. This is simply because the inputs in the previous waves were also subjected to a joint check with the respondent – with no contradiction. Following this logic, it is only possible to contradict the part of the episode that was recorded in the current survey wave, not the entire episode. For the data structure, this means that the information already collected and stored in a data row for the current part of the episode (which was contradicted in the check module) is still in the dataset, but is marked in the variable `spms` with the code -20 as "episode revoked in check module". With respect to harmonization, the contradiction is taken into account by filling the harmonized episode only with values from the subspells not marked as contradicted. That is, only those subspells not contradicted are included in the harmonized spell. The end date of the respective episode is set to the interview date of the survey wave in which the last uncontradicted information for this episode was recorded.

Last but not least: In the harmonized episodes, the occupational information is newly coded based on the summarized information. Therefore, it is possible that there are differences in the values of these generated variables between subspells and the harmonized episode. For example, it may happen that a self-employed activity is reported and additional questions are asked about it, such as the professional position, the presence of a management function, and so on. In subsequent waves, the professional episode of self-employment continues, but the function has changed with the hiring of a salaried employee. This current information is transferred to the harmonized spell. As a result, the first subspell shows a self-employed person without a leading function and the harmonized spell shows a self-employed person with a leading function. Accordingly, the occupational information is recoded in the harmonized spell.

Handling of harmonized episodes

Data users can and must decide for themselves whether to use the harmonized episodes for their data analysis or to consider the information from the separate subspells that reflect changes in the characteristics of an episode over time. Both pieces of information are available in the spell datasets.

If the harmonized episodes are to be used – including episodes that consist of only one subspell and therefore did not need to be harmonized – it is sufficient to select all data rows with the value 0 in the variable `subspell`.

```
keep if subspell==0
```

After that, all episodes should be excluded that were contradicted in the check module (variable `spms=-20`) and at the same time do not belong to the harmonized episodes (variable `spext=0`).[12] As described above, this step is already included in the process of harmonizing episodes.

If, on the other hand, one does **not** want to use the harmonized episodes but the original subspells, then all data rows must be deleted where the variable `subspell` has the value 0 and at the same time the variable `spext` has the value 1. After that, all sub-episodes must be excluded as well, which were contradicted in the check module (variable `spms=-20`).

```
drop if subspell==0 & spext==1
drop if spms==-20
```

4.5 Data files

In the following section, every data file of this Starting Cohort is explained in a subsection, including a data snapshot and an example of data usage (in Stata). The examples are written so that everyone knowing Stata should easily understand it. Also, you do not need additional ado files installed, although you are highly advised to use the `NEPStools` (see section 1.6).

To ease your understanding of the relationship of those files, figure 20 provides an overview. The edges in this graph symbolize how a data file may be linked to other files. This is not meant to document every possible data link you could do but rather tries to give you an idea which data files relate most. By clicking on a node, you get directed to this data file's explanatory page.

You need to set the following globals for the Stata examples to work. Just adapt and copy the lines below to the top of the syntax files or execute them in your Stata command line before running the syntax:

```
** Starting Cohort
global cohort SC3
** version of this Scientific Use File
global version 12-0-0
** path where the data can be found on your local computer
global datapath Z:/Data/${cohort}/${version}
```

12 The variable `spgen` also indicates whether an episode was originally reported as finished (`spgen=0`) or whether it is a harmonized (generated) episode (`spgen=1`).

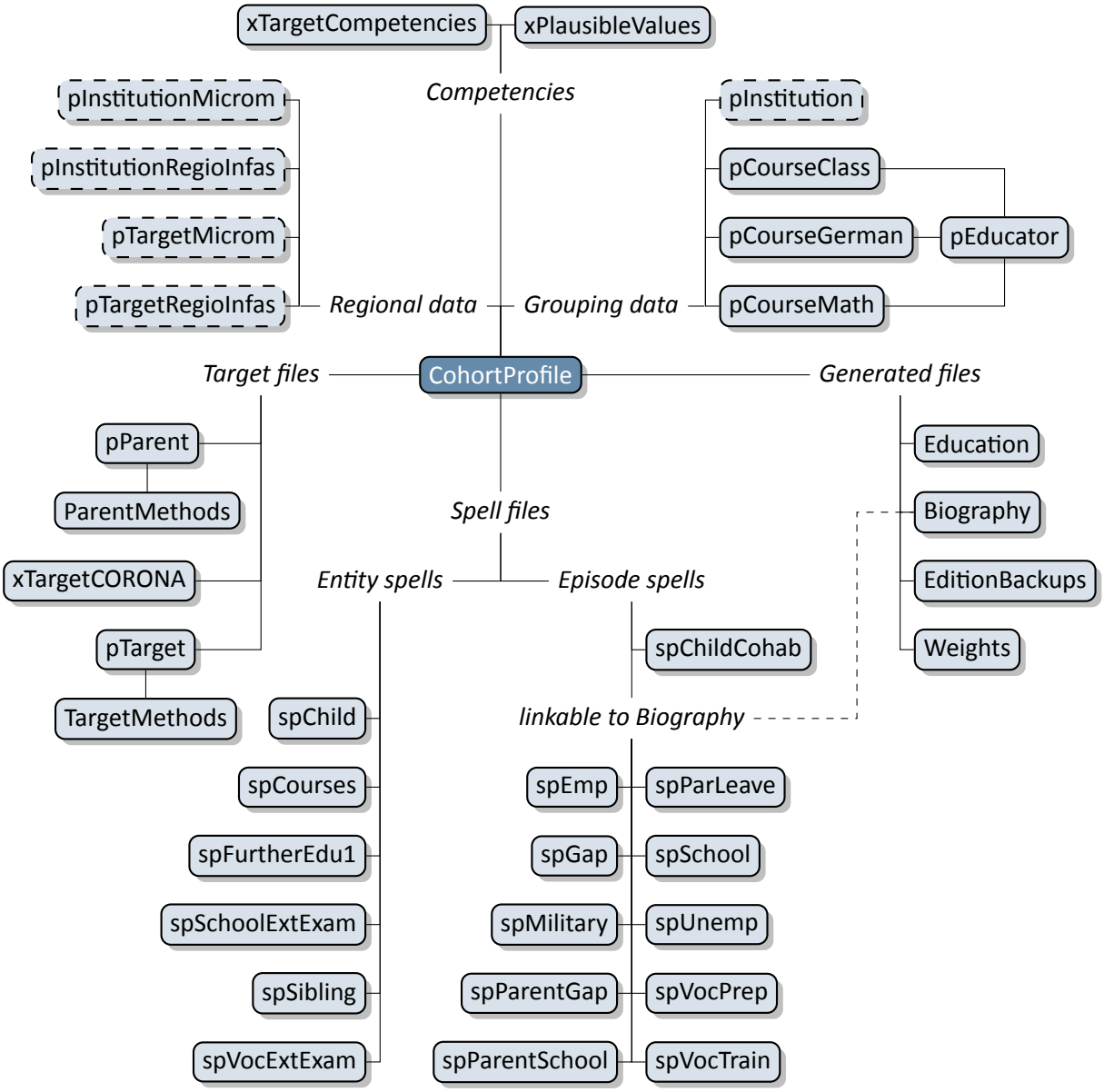

Figure 20: Graphical overview of all data files. Each node represents one data file. Relations are indicated by connection lines. Files with a dashed border are not available in the Download version of the Scientific Use File. Click on a data file to get more information.

4.5.1 Biography

« go back to overview

Description

Integrated and edited life course data

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t splink

Other ID variables useful for linkage

wave sptype

Number of variables / number of rows in file

10 / 34,817

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
wave	Wave
sptype	Spell type
startm	Episode start (month)
starty	Episode start (year)
endm	Episode end (month)
endy	Episode end (year)
spms	Type of event
splast	Episode is ongoing

Exemplary data snapshot

ID_t	splink	wave	sptype	starty	endy
4006492	220004	8	22	2012	2015
4010364	260001	10	26	2016	2018
4010651	260004	11	26	2018	2018
4027809	220001	8	22	2005	2010
4039764	220002	10	22	2010	2018

The file `Biography` serves to facilitate the analysis of complex life course data collected both retrospectively and prospectively. The dataset puts together harmonized episodes with educational and employment relevance from the following duration spell files: `spSchool`, `spVocPrep`, `spVocTrain`, `spMilitary`, `spEmp`, `spUnemp`, `spParLeave`, and `spGap`. The variable `sptype` is provided to identify the source of each episode.

In contrast to the “raw” biographical data from each of the module-specific spell modules, the `Biography` file provides more consistent life course data that has been additionally checked and edited. In particular, inconsistencies in the individual life course data were identified and corrected during the interview with the help of a “check module”. Corrected times are stored in the duration spell files as `_g1` variables. For example, the variable `ts2311y_g1` in `spEmp` contains the starting date of an employment episode which was corrected within the check module. Such corrected times form the basis for further adjustments that are implemented in

the data editing process for Biography. Essentially, the following measures are taken to ensure the integrity of the life course data in this file:

- All subspells have been removed, i. e., Biography contains only completed, harmonized, or right-censored episodes (subspell=0).
- Episodes revoked by respondents during the interview or in the subsequent survey wave (see section 4.4.2) are deleted, unless the episode was re-recorded in the current wave. Revoked episodes are included in the original spell files and can be identified there with the corresponding marker variables (spms or disagint).
- Start and end dates of episodes are smoothed and corrected, i. e., overlaps of one month and more between adjacent episodes have been resolved.
- Gaps between adjacent episodes that do not exceed two months are closed; gaps of more than two months are defined as specific gap episodes (edition gaps) within the Biography file.

Due to the additional editing steps and the compilation of spells from different biographical modules, it is recommended to use the Biography dataset as a starting point for life course analyses.

Stata 1: Working with Biography (find R example here)

```
** open the data file
use ${datapath}/${cohort}_Biography_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** check out which spell modules you can merge to this file
tab sptype

** check that you will need splink to merge information
** from other modules to this file
** (command gives no result, which means approval)
isid ID_t splink
```

4.5.2 CohortProfile

« go back to overview

Description

Paradata on the cohort's panel sample

File structure

long format: 1 row = 1 respondent in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_cc ID_cg ID_cm ID_tg ID_i

Number of variables / number of rows in file

48 / 95,394

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
cohort	NEPS Starting Cohort
tx80220	Participation/drop-out status
tx80521	Data available: survey target person
tx80522	Data available: competence test target person
tx8610m	Competence testing Target person: survey month 1
tx8610y	Competence testing Target person: survey year 1
tx80524	Data available: institution
tx80107	Sample: First participation in wave

Exemplary data snapshot

ID_t	wave	tx80220	tx80521	tx80522	tx8610y	tx80524
4007383	3	Participation	yes	yes	2012	yes
4009500	1	Participation	yes	yes	2010	yes
4009890	5	Participation	yes	yes	2014	yes
4010738	2	Participation	yes	yes	2011	yes
4041003	5	Participation	yes	yes	2014	yes

The file `CohortProfile` contains all target persons of the panel sample. These are all targets with an initial agreement to participation. For each respondent in each wave, the `CohortProfile` contains all ID variables related to this person (from class ID `ID_cc` to school ID `ID_i`), but also meta information like various variables indicating participation e. g., (`tx80220`), current grade (`tx80234`), or availability of specific data (e. g., `tx80522`). In addition, there are variables of the dates when the competence tests (`tx8610/tx8611`) took place.

In general, we strongly recommend using this file as a starting point for any analysis!

Stata 2: Working with CohortProfile

```
** open the data file
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** how many different respondents are there?
distinct ID_t

** as you can see, in this file there is an entry for every
** respondent in each wave
tab wave

** check participation status by wave
tab wave tx80220
```

4.5.3 EditionBackups

« go back to overview

Description

Backup of original data that were modified during the data edition process

File structure

long format: 1 row = 1 changed value of a variable in a data file

ID variables needed to identify a single row

dataset varname ID_t wave splink subspell partner child

Other ID variables useful for linkage

mergevars

Number of variables / number of rows in file

14 / 1,006

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
dataset	Dataset name
varname	Variable name
mergevars	ID-Variables for merging
sourcevalue_num	Original value (if numeric)
editvalue_num	New value (if numeric)
sourcevalue_str	Original value (if string)
editvalue_str	New value (if string)

Exemplary data snapshot

ID_t	wave	dataset	varname	mergevars	sourcevalue_num	editvalue_num
4006697	.	spParentSchool	p723080	ID_t splink subspell	14.00	5.00
4007193	.	spVocTrain	ts15201	ID_t splink subspell	17.00	3.00
4008886	10	pTarget	t731314	ID_t wave tx20100	21.00	16.00
4010017	1	pParent	p731818	ID_t wave	5.00	4.00
4010017	1	pParent	p731868	ID_t wave	5.00	4.00

The dataset `EditionBackups` consists of single values that have been changed or modified in the data edition process. These single values can potentially originate from all other datasets. `EditionBackups` contains both the original and the changed value of a particular variable in a particular data file (i. e., one change or edition per row). The following variables are provided for each change:

- `varname` and `dataset` specify the name of the variable affected by an edition and the respective data file
- `mergevars` lists the identifier variables that are required to merge the information back to the respective data file

- `sourcevalue_[num/str]` contains the original, unaltered value; variables with the suffix `_num` refer to values from numeric variables and variables with the suffix `_str` refer to values from string variables (if the variable is numeric, `_str` is used to store the value label for this value instead)
- `editvalue_[num/str]` contains the result of the modification, i. e. the value into which the original value was changed; these values correspond exactly to the values in the respective data file (again, there is a version for both numeric and string variables - or the label).
- `ID_t`, `wave`, ... are the different identifier variables needed to merge the orginal values to the respective data files

Stata 3: Working with EditionBackups (find R example here)

```
** In this example, we want to restore the original values in the variable
** t731353 (Highest professional qualification Father) of datafile pTarget

** open the datafile
use ${datapath}/${cohort}_EditionBackups_D_${version}.dta, clear

** only keep rows containing data of the aforesaid variable
keep if dataset=="pTarget" & varname=="t731353"

** check which variables we need for merging
tab mergevars

** then keep the merging variables and the variable with
** the original values (for cross-checking, we also keep the
** variable editvalue, which contains the values found in pTarget)
keep ID_t wave tx20100 sourcevalue_num editvalue_num

** rename the variables to emphasize affiliation
rename sourcevalue_num t731353_source
rename editvalue_num t731353_edit

** temporary save this data extract
tempfile edition
save `edition'

** open pTarget
use ${datapath}/${cohort}_pTarget_D_${version}.dta, clear

 ** add the above data
merge 1:1 ID_t wave tx20100 using `edition', keep(master match)

** check all edition made
list ID_t wave t731353* if _merge==3

** replace the variable in the datafile with its original value
replace t731353=t731353_source if _merge==3
```

4.5.4 Education

« go back to overview

Description

Generated: upward transitions in educational careers

File structure

spell format: 1 row = 1 event (episode) of 1 respondent

ID variables needed to identify a single row

ID_t number

Other ID variables useful for linkage

splink exam tx28100

Number of variables / number of rows in file

11 / 11,824

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
number	Sort number
datem	valid since (month)
datey	valid since (year)
tx28101	Recent CASMIN
tx28102	years of education = f(CASMIN)
tx28103	Recent ISCED-97
tx28109	Change in educational classification
splink	Link for spell merging
exam	Exam number
tx28100	Source of information of educational qualification

Exemplary data snapshot

ID_t	number	datey	tx28101	tx28102	tx28103	splink	tx28100
4039733	2	2018	5	13	3	220002	22
4027704	2	2016	1	9	1	220002	22
4037369	3	2018	5	13	3	220003	22
4040725	2	2018	5	13	3	220002	22
4007108	2	2016	3	10	2	220003	22

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). Also, data from `spVocExtExam` and `spSchoolExtExam` have been integrated. Three measures of educational attainment are available: CASMIN (variable `tx28101`), ISCED-97 (`tx28103`), and years of education (`tx28102`; derived from CASMIN). You can easily merge data from the original spells to Education using the variable splink. The file stores transitions in a long event time 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 considered 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 are captured by only one of these classifications.

Stata 4: Working with Education (find R example here)

```
** we want to merge the school type from spSchool to this data file.
** For this to work, we first have to prepare spSchool and keep only
** harmonized episodes (subspell==0)
use ${datapath}/SC3_spSchool_D_${version}.dta, clear
label language en
keep if subspell==0
tempfile temp
save `temp'

** now, open the Education data file
use ${datapath}/SC3_Education_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** check out which spell modules you can merge to this file
tab tx28100

** only keep school episodes
keep if tx28100==22

** check that you will need splink to merge information
** from other modules to this file
** (command gives no result, which means approval)
isid ID_t splink, miss

** merge the previously generated temporary data file
merge 1:1 ID_t splink using `temp', keep(master match) keepusing(ts11204)

** see that this only added information to the rows corresponding to spSchool
tab tx28100 _merge
```

4.5.5 ParentMethods

« go back to overview

Description

Paradata from the parents CATI interview

File structure

long format: 1 row = 1 parent in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_int

Number of variables / number of rows in file

31 / 26,012

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
px80200	Interview: number of all contact attempts
px80209	Interview: length of interview (minutes)
px80212	Interview: change of contact person to previous wave
px80214	Interview: relationship of respondent to the target child
ID_int	Interviewer: ID
px80301	Interviewer: gender
px80302	Interviewer: age group
px80207	Interview: response code differentiated
px80400	Willingness: panel participation

Exemplary data snapshot

ID_t	wave	px80209	ID_int	px80301	px80302
4006344	1	28.38333	1148	[m] male	up to 29 years
4008230	6	36.31667	1050	[m] male	30-49 years
4009059	6	27.79445	1159	[m] male	30-49 years
4009498	4	29.70000	2164	[m] male	up to 29 years
4025893	2	29.45000	1105	[w] female	30-49 years

This dataset offers a variety of information on the data collection during the interview with the parent, e. g., gender (px80301) and age (px80302) of the interviewer; survey mode (px80202); interview duration (px80209); response code (px80207).

Importantly this file contains all contacted parents, whether an interview was realized or not (see variable px80207 for more details). Thus, ParentMethods includes more cases than the data file pParent.

Stata 5: Working with ParentMethods (find R example here)

```
** open the data file
use ${datapath}/SC3_ParentMethods_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** check out response code by wave
tab wave px80207

** how many different interviewers did CATI surveys?
distinct ID_int

** get an overview on the count of contact attempts
summarize px80200
```

4.5.6 pCourseClass

« go back to overview

Description

Data about the class background

File structure

long format: 1 row = 1 school class in 1 wave

ID variables needed to identify a single row

ID_cc wave ex20100

Other ID variables useful for linkage

ID_i ID_e

Number of variables / number of rows in file

84 / 2,739

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_cc	Course-ID: Grade
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
e451010	Class: number of students with a migrant background (approximately)
e79202a_D	Students with at least one parent with a higher education degree (in %)
e227400_R	Class: number of female students
e227401_R	Class: number of male students
e22740a	Class: teacher assessment: interest
e22941e	Visualization aids, beamer
e22941f	Visualization aids, computer
e190013	no experience with special needs
ef0001a	Help write applications
ef0001c	Feeling of beeing responsible
ef0001e	Discuss individual opportunities

Exemplary data snapshot

ID_cc	wave	ID_e	ID_i	e227400_R	e227401_R	e22941f
1000235101	3	1010899	1000235	13	6	1
10005541001	7	1019859	1000554	7	11	1
1002413101	5	1013356	1002413	13	10	1
1002436102	5	1019233	1002436	9	10	2
10010871004	7	1011242	1001087	13	15	2

This data file contains all the information surveyed from the class teacher about the school classes. This is for example, besides others, the number/percentage of girls (`e227400_D`), boys (`e227401_D`), students in total (`e227400_g1D`), and students with a migration background (percentage in `e451000_D`), size of classroom (`e229400_D`), or the condition of the classroom (e. g. brightness `e22940a`). The educator reporting this information can be identified via `ID_e`.

In some cases, more than one educator reported information about a single class, although this was not intended by the survey design. In such cases, we made a suggestion which data to use in variable `ex20100`.[13]

Please note that in order to merge this data file to others, you first have to remove or aggregate duplicate classes (see example for how to do this with variable `ex20100`).

Stata 6: Working with pCourseClass (find R example here)

```
** open the data file
use ${datapath}/SC3_pCourseClass_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** only keep recommended data rows
keep if ex20100==1

** check uniqueness of key variables
isid ID_cc wave

** reduce data file to some information
keep ID_cc wave e22940b e227400_g1D

** save file for later use
tempfile pcc
save `pcc'

** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** add the class information to this file. Note that
** merge is m:1, as multiple students belong to the same class
merge m:1 ID_cc wave using `pcc', assert(master match) nogen

** crosstab some variables
tab t723080_g1 e22940b
```

13 The data row with the least missing values is being suggested.

4.5.7 pCourseGerman

« go back to overview

Description

Data about the everyday life in german class

File structure

long format: 1 row = 1 german class in 1 wave

ID variables needed to identify a single row

ID_cg wave ex20100

Other ID variables useful for linkage

ID_i ID_e

Number of variables / number of rows in file

168 / 2,524

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_cg	Course-ID: German
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
ex20100	Data line recommended for linkage
e10029a	Collaboration: reference group
e10037a	Collaboration: meetings with concrete results
ed0001h_R	Lessons German (number)
ed0001m_R	Lessons German (minutes)
ed0004a	Social methods - student groups
e538021	Weekly time - discussing homework
e538023	Weekly time - tasks with assistance
e538026	Weekly time - tests, quizzes or question games
ed00100	Time (week) for spelling
ed0012m	Spelling homework per week (minutes)

Exemplary data snapshot

ID_cg	wave	ex20100	e10037a	e538021	e538023	e538026
1000955103	2	1	3	10	10	5
1000336103	3	1	2	10	10	5
1002431301	4	1	3	10	15	10
1000362101	2	1	1	20	5	5
1000235102	3	1	3	10	5	15

This data file contains all the information surveyed from the German teachers about the German classes. This is primarily the percentage of time spent each week on different activities during class, such as time spent for discussing homework (`e538021`), listening to teacher presentations (`e538022`), taking tests (`e538026`), and others. Here, as well, the educator reporting this information can be identified via `ID_e`.

In some cases, more than one educator reported information about a single class, although this was not intended by the survey design. In such cases, we made a suggestion which data to use

in variable ex20100.[14]

Please note that in order to merge this data file to others, you first have to remove or aggregate duplicate classes (see example for how to do this with variable ex20100).

Stata 7: Working with pCourseGerman (find R example here)

```
** open the data file
use ${datapath}/SC3_pCourseGerman_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** only keep recommended data rows
keep if ex20100==1

** check uniqueness of key variables
isid ID_cg wave

** reduce data file to some information
keep ID_cg wave e538021

** save file for later use
tempfile pcg
save `pcg'

** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** add the class information to this file. Note that
** merge is m:1, as multiple students belong to the same class
merge m:1 ID_cg wave using `pcg', assert(master match) nogen

** crosstab some variables
tab e538021 tx80501
```

14 The data row with the least missing values is being suggested.

4.5.8 pCourseMath

« go back to overview

Description

Data about the everyday life in math class

File structure

long format: 1 row = 1 math class in 1 wave

ID variables needed to identify a single row

ID_cm wave ex20100

Other ID variables useful for linkage

ID_i ID_e

Number of variables / number of rows in file

107 / 2,737

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_cm	Course-ID: Mathematics
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
e10029b	Collaboration: reference group
ed0025h_R	Lessons Mathematics (number)
ed0025m_R	Lessons Mathematics (minutes)
ed0028j	Social methods - explaining
ed0030b	Forms of teaching - time to solve
ed0033a	Student groups - demands
e538011	Weekly time - discussing homework
e538012	Weekly time - teacher presentations
e538016	Weekly time - tests, quizzes or question games
e538017	Weekly time - classroom management
e538018	Weekly time - other student activities

Exemplary data snapshot

ID_cm	wave	ID_e	ID_i	e538011	e538012	e538016
1000329102	2	1003393	1000329	15	10	5
1002383203	5	1019333	1002383	10	20	10
1000262102	3	1003205	1000262	10	15	8
1000632102	2	1003203	1000632	25	5	5
1002445202	3	1010701	1002445	5	10	5

This data file contains all the information surveyed from the math teachers about the math classes. This is primarily the percentage of time spent each week on different activities during class, such as time spent for discussing homework (e538011), listening to teacher presentations (e538012), taking tests (e538016), and others. Here, as well, the educator reporting this information can be identified via ID_e.

In some cases, more than one educator reported information about a single class, although this was not intended by the survey design. In such cases, we made a suggestion which data to use in variable `ex20100`.[15]

Please note that in order to merge this data file to others, you first have to remove or aggregate duplicate classes (see example for how to do this with variable `ex20100`).

Stata 8: Working with pCourseMath (find R example here)

```
** open the data file
use ${datapath}/SC3_pCourseMath_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** only keep recommended data rows
keep if ex20100==1

** check uniqueness of key variables
isid ID_cm wave

** reduce data file to some information
keep ID_cm wave e538011

** save file for later use
tempfile pcm
save `pcm'

** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** add the class information to this file. Note that
** merge is m:1, as multiple students belong to the same class
merge m:1 ID_cm wave using `pcm', assert(master match) nogen

** crosstab some variables
tab e538011 tx80501
```

15 The data row with the least missing values is being suggested.

4.5.9 pEducator

« go back to overview

Description

Personal information about the teachers

File structure

long format: 1 row = 1 educator in 1 wave

ID variables needed to identify a single row

ID_e wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

302 / 4,439

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_e	ID teacher/educator
wave	Wave
e400000	Migrant background teacher
e41100a_g1	Mother tongue (number references)
e537010	Teaching experience before college
e537090	Teaching degree course
e537150_R	Year of state examination
e762110	Gender
e537180	Grade First state examination
e537190	Second state examination
e537210	Grade Second state examination
e76212y_R	Year of birth

Exemplary data snapshot

ID_e	wave	e400000	e41100a_g1	e537090	e537150_R	e762110
1019982	7	3	1	1	1985	[m] male
1019404	5	3	1	1	1996	[w] female
1020072	7	3	1	1	1999	[w] female
1010817	3	3	1	1	1998	[m] male
1003258	1	3	1	1	1997	[w] female

Teachers were interviewed as context persons during the school survey by a PAPI questionnaire. This data is made available in file `pEducator`. The scope of information comprises various personal attributes of the educator, e. g., gender (`e762110`), year of birth (`e76212y_D`), or amount of years in occupation (`e229820_D`), as well as attitudes or sensitivities, such as stress factors (e g., `ed1009a`) or aspects of career choice (`e536031`).

This file contains all educators from the sample, no matter if they were teaching class, German, or math classes. Note that there is no direct link between students and teachers made available. This is due the following reasons:

- By study design, one student could have up to three teachers (class, German, math).
- Due to missing detail in instructions, in some cases more than on class/German/math teacher answered the survey (see variable `ex20100` in, e. g., `pCourseClass`).

- There is no natural relationship in the survey design between teachers and students. Teachers have only been interviewed about themselves, as well as about the class. There were no questions asked to the teachers about individual students.

To map those facts to the data, the educator was only attached to the specific classes (i. e., `pCourseClass`, `pCourseGerman`, and `pCourseMath`), and not to the students directly (i. e., there is no variable `ID_e` in `CohortProfile`). See the example below on how you are supposed to use the data from `pEducator`.

Stata 9: Working with pEducator (find R example here)

```
** Goal: Have class-teachers gender available in CohortProfile
** (i.e., on student level). Walkthrough:
** a) collect data from pEducator and simplify structure
** b) merge this data to pCourseClass, retaining easy structure
** c) merge combined data to CohortProfile

** a)
** open data from pEducator file
use ID_e wave e762110 using ${datapath}/SC3_pEducator_D_${version}.dta, clear
label language en

** this data file is still in panel logic (i.e., one row per wave), although
** the data itself is time-invariant! To ease later merging, we reduce
** complexity of this file by restructuring to a cross-sectional format.

** remove missing rows
nepsmiss e762110
drop if missing(e762110)

** remove duplicates; in case of discrepancy, keep data from first wave
sort ID_e wave
drop wave
duplicates drop ID_e, force

** save for later merge
tempfile ped
save `ped'

** b)
** open class data file pCourseClass
use ID_cc wave ID_e ex20100 using ${datapath}/SC3_pCourseClass_D_${version}.dta,
  clear
label language en

** only keep recommended data rows
keep if ex20100==1
drop ex20100

** merge pEducator-data
merge m:1 ID_e using `ped', keep(master match) nogen

** save for later merge
```

```
tempfile pcc
save `pcc'

** c)
** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear
label language en

** merge the data
merge m:1 ID_cc wave using `pcc', assert(master match) nogen

** crosstab gender of child to gender of teacher
tab tx80501 e762110
```

4.5.10 pInstitution

« go back to overview

Description

Context data collected from the head of institution about the school

File structure

long format: 1 row = 1 school in 1 wave

ID variables needed to identify a single row

ID_i wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

422 / 1,335

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_i	Institution ID
wave	Wave
h190011	Number Students with special needs
h22202a	School: quality: complete school mission statement
h22202h	School: quality: class tests
h227000	School: teaching staff: number of teachers
h229000	School: administration
h535010	Schools within a radius of 10 km
h22900a	School: structure: half-day school
h229010	School: grade levels, minimum
h229011	School: grade levels, maximum
h535021	Intensity of competition
h535023	Existence at risk

Exemplary data snapshot

ID_i	wave	h190011	h22202a	h227000	h229000	h535010
1000270	5	3	1	60	1	1
1000641	5	2	1	98	1	1
1000550	5	2	1	72	1	8
1000291	5	5	1	45	1	5
1000651	5	35	1	31	1	4

Data about the school itself has been surveyed from the principal of the institution via PAPI mode. The resulting data file `pInstitution`) contains this information, including data like the number of classes (5th grade in `h229020`, 7th grade in `h229023`, 9th grade in `h229021`), the total number of students (`h227100`), as well as background on the infrastructure (e. g., schools within a radius of 10km in `h535010`), and more.

Please note that this datafile is only available in RemoteNEPS!

Stata 10: Working with pInstitution (find R example here)

```
** open the CohortProfile
use ${datapath}/SC3_CohortProfile_R_${version}.dta, clear

** merge the size of the school to CohortProfile using school ID
merge m:1 ID_i wave using ${datapath}/SC3_pInstitution_R_${version}.dta, ///
  keepusing(h227100) nogen assert(master match)

** change language to english (defaults to german)
label language en

**cluster the children according to the quantiles of the institution size
xtile size = h227100, nq(5)

tab size
```

4.5.11 pInstitutionMicrom

« go back to overview

Description

regional data about the geographical area of institution

File structure

panel format: 1 row = 1 regional level in 1 wave of 1 institution

ID variables needed to identify a single row

ID_i wave regio

Other ID variables useful for linkage

ID_regio

Number of variables / number of rows in file

188 / 3,776

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_i	Institution ID
wave	Wave
regio	Indicator for enrichment level
ID_regio	System-free ID of enrichment level
mso_k_ausland	Share foreigners
mso_k_familie	Family structure
mbe_k_haustyp	Type of house
mgm_k_dom	Dominant microm geo milieu®
mgs_k_dom	Dominant geo-submilieu
mmo_k_volumen	Move volume
mpi_k_dichte	Car density
mas_k_berufsuv	Occupational disability insurance
mas_k_krankzuv	Additional health insurance
mlt_k_primlt	Primary Limbic Type
kkr_w_summe	Total purchasing power in euros

Exemplary data snapshot

ID_i	wave	regio	ID_regio	mso_k_ausland	mbe_k_haustyp	mpi_k_dichte
1000511	3	1	140601	7	7	4
1000628	3	1	102550	7	2	2
1000614	3	1	127245	5	2	5
1002387	3	1	115319	5	2	7
1000500	3	1	120854	7	7	5

The data file `pInstitutionMicrom` is only available **Onsite**. You cannot work with this file having only access to the Download or RemoteNEPS SUF.

It contains some regional details on the geographical area of the institution on five different regional levels: house area, road section, postal code, postal code 8, municipality.

All those levels are available for every institution and every wave. There is a lot of regional information in this file, including percentage of foreigners, unemployment rate, family structure, milieu types, car type/density, insurances, only to name a few. To clarify this, those details

are **not** about the respondents or the institutions but about the regional level (e. g., the unemployment rate is not the rate at the institution but the rate in the municipality the institution resides). Please be aware that there is a complete documentation about this data file that not only lists all variables but also has a description of the background. See section 1.2 on how to find this document.

Stata 11: Working with pInstitutionMicrom (find R example here)

```
** open Microm data file. Note that this data file is only available OnSite!
use ${datapath}/${cohort}_pInstitutionMicrom_0_${version}.dta, clear

** additionally to ID_i and wave, line identification in this file is done
** via variable regio, denoting the regional level of information
isid ID_i wave regio

** tabulating wave against regio shows availability of all levels
** in wave 5, but only the most detailled level available
** in wave 1 and 3 (usually housing level)
tab wave regio

** only keep housing level
keep if regio==1

** save to temporary file
tempfile regio
save `regio'

** now you can enhance CohortProfile with regional data
use ${datapath}/${cohort}_CohortProfile_0_${version}.dta, clear
label language en

merge m:1 ID_i wave using `regio'
```

4.5.12 pInstitutionRegioInfas

« go back to overview

Description

regional data about the geographical area of institution

File structure

panel format: 1 row = 1 regional level of 1 institution

ID variables needed to identify a single row

ID_i regio

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

68 / 1,160

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_i	Institution ID
regio	Regional level
tx44288	Share residents 0-14 years (in %)
tx44289	Share residents 15-24 years (in %)
tx44294	Purchasing power per resident (EUR)
tx44298	Companies in total per km² (trade indicator)
tx44302	Share retail (in %)
tx44001	Residents per household
tx44318	Share single-person households (in %)
tx44242	Type of residential area
tx44312	Share agriculture (in %)
tx44354	Share residential type post-war apartment complex (in %)

Exemplary data snapshot

ID_i	regio	tx44288	tx44294	tx44298	tx44001	tx44318
1000388	2	12.6	18061	49.86	2.02	36.93
1000655	3	13.0	18119	491.86	1.83	44.30
1000262	3	15.2	17264	570.27	1.95	41.10
1000371	1	11.9	17886	192.09	1.77	51.28
1000335	2	15.4	17176	10.43	2.01	35.82

The data file `pInstitutionRegioInfas` is only available **Onsite**. You cannot work with this file having only access to the Download or RemoteNEPS SUF.

It contains some regional details on the geographical area of the institution on four different regional levels: street section, quarter, postal code, and municipality. All those levels are available for wave 1 only.

There is a lot of regional information in this file, including purchasing power per resident in EUR (`tx44294`), companies in total per km^2 (`tx44298`), residents per household (`tx44001`), and so on. As in `pTargetMicrom`, those details are **not** about the respondents but about the regional level (e. g., the unemployment rate is not the rate at the institution but the rate in this municipality). Please be aware that there is a complete documentation about this data file that

not only lists all variables but also has a description of the background. See section 1.2 on how to find this document.

Stata 12: Working with pInstitutionRegioInfas (find R example here)

```
** open data file. Note that this data file is only available OnSite!
use ${datapath}/${cohort}_pInstitutionRegioInfas_0_${version}.dta, clear
label language en

** identification in this file is done
** via variable regio, denoting the regional level of information
isid ID_i regio

** existing regional levels are:
tab regio

** only keep housing level
keep if regio==1

** save to temporary file
tempfile regio
save `regio'

** now you can enhance CohortProfile with regional data
use ${datapath}/${cohort}_CohortProfile_0_${version}.dta, clear
label language en

merge m:1 ID_i wave using `regio'
```

4.5.13 pParent

« go back to overview

Description

Data surveyed from parents

File structure

long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

733 / 20,063

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
p406010_g1R	Country of birth of target child
p731905	Professional status Respondent
p731955	Professional status Partner
p731701	Relationship to target child
p741001	Size of household
p510005	monthly household income, open
p400500_g1	Generation status
p743040	Child in household
p731905	Professional status Respondent
p751001_g2R	Place of residence (federal state)
p34009d	Participation in high culture: theater
p73170y	Date of birth respondent: year
p401100	Citizenship Respondent
p731116	Gender Partner

Exemplary data snapshot

ID_t	wave	p731905	p731955	p741001	p400500_g1	p743040
4007560	1	2	2	4	6	yes
4008497	1	2	2	4	6	yes
4010544	1	5	2	3	5	yes
4027730	1	5	5	3	3	yes
4040438	3	5	2	3	10	yes

Parents' interviews from both the CATI and the CAPI module are stored in the data file `pParent`. Various topics were recorded, ranging from personal attributes of the parent or her or his partner, e. g., occupation or respondent (`p731905`) and occupation of partner (`p731955`), to household specific matters, e. g., size of household (`p741001`), to subjects related directly to the child, e. g., size and weight of child at birth (`p529000`, `p529001`). Note that some information collected from the parents is in episode format; thus, it is not stored in data file `pParent`, but in separate spell modules (e. g., `spParentSchool`).

Stata 13: Working with pParent (find R example here)

```
** open the CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** merge occupation of parents (both respondent and partner) from pParent
merge 1:1 ID_t wave using ${datapath}/SC3_pParent_D_${version}.dta, ///
  keepusing(p731905 p731955) nogen assert(master match)

** change language to english (defaults to german)
label language en

** recode missings
nepsmiss p731905 p731955

** note that parent data is only available in certain waves
tab p731905 wave, miss

** thus, to work with this information in other waves, you
** first have to carry over the values to other rows
bysort ID_t (wave): replace p731905=p731905[_n-1] if missing(p731905)
bysort ID_t (wave): replace p731955=p731955[_n-1] if missing(p731955)

** check the distribution of parents occupation in current type of school
tab2 p731905 p731955 t723080_g1
```

4.5.14 pTarget

« go back to overview

Description

Data surveyed from the targets

File structure

long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

2,837 / 61,961

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
t34005a	Number of books
t514001	Satisfaction with life
t400000_g1D	Country of birth (Germany/abroad)
t400500_g1	Generation status
t41000a_g1	Mother tongue (number references)
t514004	Satisfaction with family life
t66002a_g1	Self concept school
t66003a_g1	Global self-esteem

Exemplary data snapshot

ID_t	wave	t514001	t400000_g1D	t400500_g1	t41000a_g1	t514004
4006666	1	10	not Germany	1	1	5
4008068	1	5	not Germany	1	2	8
4040052	3	8	not Germany	2	5	10
4040471	3	7	not Germany	2	1	10
4041134	3	3	not Germany	2	2	2

The file `pTarget` contains all topics surveyed via PAPI questionnaire from the target persons (students) themselves. A lot of items are included in this data file, ranging from demographic topics (e. g., citizenship in `t40115a_g2D`) to attitudes or expectations (e. g., first job wish in `tf00070_g1`, activities (e. g., sports in `t262000_g1`), network issues (e. g., number of friends with Abitur goal in `t32111c`), and many more.

Stata 14: Working with pTarget (find R example here)

```
** open the CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** as there are multiple instances of some IDs in a specific wave, we
** need this 'hack' to deduplicate the data during the following merge process
gen tx20100=1

** merge country of birth and generation status from pTarget
merge 1:1 ID_t wave tx20100 using ${datapath}/SC3_pTarget_D_${version}.dta, ///
  keepusing(t400500_g1 t400000_g1D) nogen

** change language to english (defaults to german)
label language en

** recode missings
nepsmiss t400500_g1 t400000_g1D

** note that parent data is only available in certain waves
tab t400000_g1D wave, miss

** thus, to work with this information in other waves, you
** first have to carry over the values to other rows
bysort ID_t (wave): replace t400000_g1D=t400000_g1D[_n-1] if missing(t400000_g1D)
bysort ID_t (wave): replace t400500_g1=t400500_g1[_n-1] if missing(t400500_g1)

** check the above alteration
tab t400000_g1D wave, miss

** check the distribution between migration and current type of school
tab t723080_g1 t400000_g1D
```

4.5.15 pTargetMicrom

« go back to overview

Description

Small-scale regional indicators on respondents' place of residence

File structure

panel format: 1 row = 1 regional level in 1 wave of 1 respondent

ID variables needed to identify a single row

ID_t wave regio

Other ID variables useful for linkage

ID_regio

Number of variables / number of rows in file

188 / 68,087

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
regio	Indicator for enrichment level
ID_regio	System-free ID of enrichment level
mso_k_ausland	Share foreigners
mso_k_familie	Family structure
mbe_k_haustyp	Type of house
mgm_k_dom	Dominant microm geo milieu®
mgs_k_dom	Dominant geo-submilieu
mmo_k_volumen	Move volume
mpi_k_dichte	Car density
mas_k_berufsuv	Occupational disability insurance
mas_k_krankzuv	Additional health insurance
mlt_k_primlt	Primary Limbic Type
kkr_w_summe	Total purchasing power in euros

Exemplary data snapshot

ID_t	wave	regio	ID_regio	mso_k_ausland	mbe_k_haustyp	mpi_k_dichte
4005926	4	1	132638	8	4	6
4008408	3	1	142246	3	5	3
4008694	4	1	127173	2	2	6
4008985	3	1	135065	7	1	8
4009777	3	1	123955	2	1	5

The data file `pTargetMicrom` is only available **Onsite**. You cannot work with this file having only access to the Download or RemoteNEPS SUF.

The data include details about the respondent's residence at five different regional levels, distinguishable by the variable `regio`: house area, street section, postal code, postal code 8, municipality. All these levels are available for each respondent and each wave.

Numerous regional indicators are provided, e. g. the percentage of foreigners, unemployment rate, family and age structure, milieu types, car type density, distribution of insurances, etc. To

clarify, this information does **not** refer to individuals, but to regional units to which respondents belong via their place of residence. Accordingly, the unemployment rate, for example, indicates the proportion of unemployed people in the population of a given region.

Please note that a separate documentation exists for this data file on the website (see section 1.2), which not only lists all variables, but also explains the background of the data.

Stata 15: Working with pTargetMicrom (find R example here)

```
** open Microm data file. Note that this data file is only available OnSite!
use ${datapath}/${cohort}_pTargetMicrom_O_${version}.dta, clear
label language en

** additionally to ID_t and wave, line identification in this file is done
** via variable regio, denoting the regional level of information
isid ID_t wave regio

** tabulating wave against regio shows availability of all levels
** in wave 5 and 7, but only the most detailled level available
** in wave 1 and 3 (usually housing level)
tab wave regio

** only keep housing level
keep if regio==1

** now you can enhance CohortProfile with regional data
merge 1:1 ID_t wave using ${datapath}/${cohort}_CohortProfile_O_${version}.dta
```

4.5.16 pTargetRegioInfas

« go back to overview

Description

Small-scale regional indicators on respondents' place of residence

File structure

panel format: 1 row = 1 regional level of 1 respondent

ID variables needed to identify a single row

ID_t regio

Other ID variables useful for linkage

none

Number of variables / number of rows in file

68 / 19,020

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
regio	Regional level
tx44288	Share residents 0-14 years (in %)
tx44289	Share residents 15-24 years (in %)
tx44294	Purchasing power per resident (EUR)
tx44298	Companies in total per km² (trade indicator)
tx44302	Share retail (in %)
tx44001	Residents per household
tx44318	Share single-person households (in %)
tx44242	Type of residential area
tx44312	Share agriculture (in %)
tx44354	Share residential type post-war apartment complex (in %)

Exemplary data snapshot

ID_t	regio	tx44294	tx44298	tx44001
4008319	2	17810	10.84	2.07
4006636	3	16438	106.90	2.09
4025932	2	19988	26.48	2.12
4009662	1	16365	62.88	1.76
4006834	2	18132	20.16	1.98

The data file `pTargetRegioInfas` is only available **Onsite**. You cannot work with this file having only access to the Download or RemoteNEPS SUF.

The data include details about the respondent's residence at four different regional levels, distinguishable by the variable `regio`: street section, quarter, postal code, and municipality. All those levels are available for wave 1 only. At this time, the address was only known for those students whose parents were willing to participate in the study (although not necessarily participated in the end). Thus, the file does not contain information for the complete sample of wave 1. The regional indicators available in this file include the purchasing power per resident in EUR (`tx44294`), the total number of companies per km^2 (`tx44298`), the average number of residents per household (`tx44001`), and so on. As in `pTargetMicrom` these data do **not** refer to the respondents themselves, but to the regional levels in which the respondents live (i. e.,

the unemployment rate, for example, indicates the proportion of unemployed people in the population of a given region such as the municipiality).

Please note that a separate documentation exists for this data file on the website (see section 1.2), which not only lists all variables, but also explains the background of the data.

Stata 16: Working with pTargetRegioInfas (find R example here)

```
** open data file. Note that this data file is only available OnSite!
use ${datapath}/${cohort}_pTargetRegioInfas_O_${version}.dta, clear
label language en

** identification in this file is done
** via variable regio, denoting the regional level of information
isid ID_t regio

** existing regional levels are:
tab regio

** only keep housing level
keep if regio==1


** save to temporary file
tempfile regio
save `regio'

** now you can enhance CohortProfile with regional data
use ${datapath}/${cohort}_CohortProfile_O_${version}.dta, clear
label language en

merge 1:1 ID_t wave using `regio'
```

4.5.17 spChild

« go back to overview

Description

information about all children of respondent

File structure

entity format: 1 row = 1 child of 1 respondent

ID variables needed to identify a single row

ID_t child subspell

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

64 / 85

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
child	Child number
subspell	Number of subspell
wave	Wave
ts3320m	Date of birth of the child (month)
ts3320y	Date of birth of the child (year)
ts33203	Gender of the child
ts33204	Biological child, adoptive or foster child
ts33209	Employment child
ts33216	Vocational training, child

Exemplary data snapshot

ID_t	child	subspell	wave	ts3320y	ts33203	ts33204
4008479	1	1	10	2018	[w] female	biological child
4008954	1	1	11	2018	[w] female	biological child
4009574	1	1	9	2017	[w] female	biological child
4039911	1	1	9	2018	[w] female	biological child
4040799	1	1	9	2016	[w] female	biological child

This module contains information on all biological, foster, and adopted children of the respondent, and any 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. 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. Spell data on cohabitation with children is stored in file `spChildCohab` and spell data on parental leaves relating to children is stored in `spParLeave`.

Stata 17: Working with spChild (find R example here)

```
** open the data file
use ${datapath}/SC3_spChild_D_${version}.dta, clear

** switch to english language
label language en

** only keep full or harmonized episodes
keep if subspell==0

** generate the total count of children for each respondent
** you can do this either by taking the maximum child number:
bysort ID_t: egen children=max(child)
** or counting the number of rows:
bysort ID_t: gen children2=_N
** which both computes the same result
assert children==children2

** recode rough values (e.g. end of year) to real months
replace ts3320m=ts3320m-20 if ts3320m>20

** compute the age of one's children today
** first, create a Stata monthly date (months since 1960m1) of the birth variables
generate birth_ym =ym(ts3320y,ts3320m)
** then, create the same for the current date
gen now_ym=mofd(date(c(current_date), "DMY"))
** the age is then easily computed
gen age=(now_ym-birth_ym)/12

summarize age
```

4.5.18 spChildCohab

« go back to overview

Description

file listing cohabitation spells with children

File structure

spell format: 1 row = 1 cohabitation time of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

child wave

Number of variables / number of rows in file

19 / 62

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
child	Child number
spell	Spell number
subspell	Number of subspell
wave	Wave
ts3331m	Start date Living together with child (month)
ts3331y	Start date Living together with child (year)
ts3332m	End date Living together with child (month)
ts3332y	End date Living together with child (year)
ts3332c	Currently living together with child

Exemplary data snapshot

ID_t	child	spell	subspell	wave	ts3331y	ts3332y
4006649	1	101	1	9	2017	2018
4006649	1	101	2	10	2017	2018
4007552	1	101	1	9	2016	2018
4007552	2	202	2	10	2017	2019
4009574	1	101	3	11	2017	2019

If a respondent lives together with children, durations are registered in `spChildCohab`. Cohabitation spells are related to children by the `child` number. Please note that those durations do not necessarily match birth and death events; rather see `spChild` for direct information on children.

Stata 18: Working with spChildCohab (find R example here)

```
** open the data file
use ${datapath}/SC3_spChildCohab_D_${version}.dta, clear

** switch to english language
label language en

** only keep full or harmonized episodes
keep if subspell==0

** recode rough values (e.g. end of year) to real months
replace ts3331m=ts3331m-20 if ts3331m>20
replace ts3332m=ts3332m-20 if ts3332m>20

** generate the following durations in months:
* a) the total duration of a cohabitation episode
gen cohab_duration = ym(ts3332y,ts3332m) - ym( ts3331y, ts3331m)
* b) the total duration a respondent lived together with specific child
bysort ID_t child (spell): egen total_duration_per_child  = total(cohab_duration)
* c) the total duration a respondent lived together with any child
bysort ID_t (child spell): egen total_duration_per_target = total(cohab_duration)

** to work with the latter information in other files, you could do
keep ID_t total_duration_per_target
duplicates drop
** which gives you a cross-sectional display of cohabitation time for every
  respondent
```

4.5.19 spCourses

« go back to overview

Description

dynamic course module

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t wave splink

Other ID variables useful for linkage

sptype course_w*

Number of variables / number of rows in file

77 / 979

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
splink	Link for spell merging
sptype	Spell type
t271001	Total duration of training courses
course_w1	1st course number
t271011_w1	Duration of the course
course_w2	2nd course number
t271011_w2	Duration of the course
course_w3	3rd course number
t27800a	Start date episode (month)
t27800b	Start date episode (year)
t27800c	End date episode (month)
t27800d	End date episode (year)

Exemplary data snapshot

ID_t	wave	splink	sptype	course_w1	course_w2	course_w3
4006581	10	270001	27	1001	1002	1003
4008645	10	250001	25	1001	1002	1003
4008658	11	250001	25	1101	1102	1103
4010079	12	250001	25	1201	1202	1203
4010810	10	250001	25	1001	1002	1003

This module comprises courses and trainings attended within the past 12 months during episodes of employment (spEmp), unemployment (spUnemp), parental leave (spParLeave), military, or civilian service (spMilitary), as well as episodes from the spGap module. The start and end dates of the spells in this module represent the original episodes (in which a course was taken) from those modules. For each of these episodes, information on up to three courses is included in wide format. spCourses comprises all spells from the past 12 months that 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. 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). Please note that this information has been integrated into data file Education. If your interest in this data is not too profound, you are best advised to use Education instead.

Stata 19: Working with spCourses (find R example here)

```
** open the data file
use ${datapath}/SC3_spCourses_D_${version}.dta, clear

** check which modules provided course information
tab sptype

** only keep courses from employment spells
keep if sptype==26

** save this datafile for later usage
tempfile courses
save `courses'

** open the employment module
use ${datapath}/SC3_spEmp_D_${version}.dta, clear

** add the temporary datafile from above;
** note that this is an m:1 merge, as there are still subspells in spEmp
merge m:1 ID_t wave splink using `courses', assert(master match) nogenerate

** you now have the spEmp datafile, enhanced with information from spCourses,
** and can proceed with this in the usual way
```

4.5.20 spEmp

« go back to overview

Description

spell data on employment episodes

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

splink wave

Number of variables / number of rows in file

121 / 12,920

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
subspell	Number of subspell
spell	Spell number
ts2311y	Start Employment episode - year
ts2312y	End Employment episode (year)
ts23410	Net income, open
ts23228	Type of required training
ts23201_g1	Professional title (KldB 1988)
ts23201_g2	Professional title (KldB 2010)
ts23201_g3	Professional title (ISCO-88)

Exemplary data snapshot

ID_t	subspell	spell	ts2311y	ts2312y	ts23410	ts23228
4009596	1	2	2019	2019	1800	3
4010490	1	2	2019	2019	1600	3
4006940	1	2	2020	2020	2000	3
4039383	1	2	2019	2019	500	3
4040231	1	1	2019	2019	1401	1

This extensive module covers all spells of regular employment, including traineeships. Information on second jobs is only collected for activities that continue up to the date of the interview. 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)

The file comprises information like net income (`ts23410`), type of required vocational training (`ts23228`), or actual working time per week (`ts23223`).

Stata 20: Working with spEmp (find R example here)

```
** open the data file
use ${datapath}/SC3_spEmp_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.21 spFurtherEdu1

« go back to overview

Description

information about additional courses

File structure

entity format: 1 row = 1 course of 1 respondent

ID variables needed to identify a single row

ID_t course

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

16 / 1,141

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
course	Course number
t271048	Course is ongoing
t271049	Termination course
t272000_g13	Content other course (course ID)
t271043	Duration of the course

Exemplary data snapshot

ID_t	wave	course	t271048
4007359	11	1101	no
4007632	10	1001	no
4009552	11	1101	no
4039411	9	901	yes, course is ongoing
4040365	9	901	no

This module contains information on further courses (also private courses) attended within the past 12 months that have not been 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). Please note that this information has been integrated into datafile `Education`. If your interest in this data is not too profound, you are best advised to use `Education` instead.

Stata 21: Working with spFurtherEdu1 (find R example here)

```
** open the datafile
use ${datapath}/SC3_spFurtherEdu1_D_${version}.dta, clear

** One row contains information for one course. The only possibility to use
** this file is to merge it to the data for this respondents wave (we use the
** CohortProfile). We have to reshape the file so one row contains one wave.
bysort ID_t wave (course): gen course_nr=_n
reshape wide course t*, i(ID_t wave) j(course_nr)

** create a temporary datafile for later merge
tempfile spfurther
save `spfurther'

** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** merge data
merge 1:1 ID_t wave using `spfurther', assert(master match) nogen

** Please note that you now have multiple variables added to CohortProfile,
** one set of variables for each course reported in spFurtherEdu1
```

4.5.22 spGap

« go back to overview

Description

reported gap episodes

File structure

spell format: 1 row = 1 gap of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

33 / 3,939

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
spell	Spell number
subspell	Number of subspell
wave	Wave
spms	Type of event
ts29901	Auxiliary variable Current gap
ts29300	Episodenmodus
ts2911m	Start date Gap (month)
ts2911y	Start date Gap (year)
ts2912m	End date Gap (month)
ts2912y	End date Gap (year)
ts2912c	Ongoing of gap episode
ts29201	Courses during gap
ts29101	Type of gap

Exemplary data snapshot

ID_t	spell	subspell	wave	ts29901	ts2911y	ts2912y
4005756	1	1	9	1	2017	2017
4007672	3	1	10	1	2018	2018
4008719	1	1	10	1	2018	2019
4010084	2	1	10	1	2019	2019
4010171	1	1	11	1	2019	2019

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 that can be distinguished by the variable ts29101 (Type of gap episode).

Stata 22: Working with spGap (find R example here)

```
** open the data file
use ${datapath}/SC3_spGap_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.23 spMilitary

« go back to overview

Description

military / civilian service and voluntary gap years

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

26 / 1,483

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
subspell	Number of subspell
spell	Spell number
wave	Wave
ts21201	Type of military service episode
ts2111m	Start Military service episode - month
ts2111y	Start Military service episode - year
ts2112m	End Military service episode - month
ts2112y	End Military service episode - year
ts21202	Attendance of training courses/courses during military service

Exemplary data snapshot

ID_t	splink	subspell	spell	wave	ts2111y	ts2112y
4006067	250001	2	1	12	2018	2019
4006951	250001	1	1	10	2018	2018
4009189	250001	1	1	10	2018	2018
4009442	250001	2	1	12	2019	2020
4039691	250001	2	1	11	2018	2019

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 found in the employment module `spEmp`.

Stata 23: Working with spMilitary (find R example here)

```
** open the data file
use ${datapath}/SC3_spMilitary_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.24 spParentGap

« go back to overview

Description

gap episodes reported by the parents

File structure

spell format: 1 row = 1 gap of 1 respondent

ID variables needed to identify a single row

ID_t spell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

17 / 262

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
spell	Spell number
wave	Wave
ps29101	Type of gap episode
ps2911m	Start date Gap
ps2911y	Start date Gap
ps2912m	End date Gap
ps2912y	End date Gap
ps2912c	Ongoing of gap episode
ps2911y_g1	Check module: start date (year), corrected
spms	Check module: spell type

Exemplary data snapshot

ID_t	wave	ps2911m	ps2911y	ps2912m	ps2912y	ps2912c
4005799	4	9	2013	2	2014	1
4006629	1	2	2011	2	2011	1
4006840	4	8	2013	1	2014	1
4007966	1	12	2010	2	2011	1
4026206	2	4	2009	5	2012	1

Analogue to the identification of gaps in the individiual life courses included in the datafile `spGap`, the datafile `spParentGap` contains those gaps of the target persons **reported by the parents during the parent CATI**. Note that these are not gaps in the lifecourse of the parent, but of the children! The spells in this file refer to different types of gaps that can be distinguished by the variable `ps29101` (Type of gap episode).

Stata 24: Working with spParentGap (find R example here)

```
** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using ///
        ${datapath}/SC3_spParentGap_D_${version}.dta, keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.25 spParentSchool

« go back to overview

Description

general schooling history reported by the parents

File structure

spell format: 1 row = 1 school episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

41 / 32,825

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
subspell	Number of subspell
spell	Spell number
wave	Wave
p723020	School attendance in Germany
p723180	School authority
p72302m	End date school episode (month)
p72302y	End date school episode (year)
p723080	School type
p723120	Reason end of school episode
p723130	Reason school change
p723140	Reason school interruption

Exemplary data snapshot

ID_t	subspell	spell	wave	p723020	p72302m	p72302y	p723080
4006344	1	2	1	1	1	2011	5
4006840	1	2	1	1	2	2011	8
4009578	1	3	1	1	2	2011	5
4009848	1	2	1	1	3	2011	8
4041707	1	3	3	1	5	2013	8

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. The data in this file is the school history reported by the parent during the parent CATI. See file `spSchool` for the school history reported by the target herself or himself.

Stata 25: Working with spParentSchool (find R example here)

```
** open the data file
use ${datapath}/SC3_spParentSchool_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.26 spParLeave

« go back to overview

Description

episodes of parental leave

File structure

spell format: 1 row = 1 parental leave episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave child splink

Number of variables / number of rows in file

36 / 32

Contains data from waves

Exemplary variables

ID_t	ID target
child	Child number
spell	Spell number
subspell	Number of subspell
wave	Wave
ts2711m	Start Parental leave (month)
ts2711y	Start Parental leave (month)
ts2712m	End parental leave (month)
ts2712y	End parental leave (year)
ts2712c	Ongoing of parental leave

Exemplary data snapshot

ID_t	child	spell	subspell	wave	ts2711y	ts2712y
4007552	2	202	1	9	2017	2018
4007552	2	202	2	10	2017	2018
4040274	1	101	2	11	2016	2018
4040274	1	101	0	11	2016	2018
4040799	1	101	0	10	2016	2019

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.

Stata 26: Working with spParLeave (find R example here)

```
** open the data file
use ${datapath}/SC3_spParLeave_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.27 spSchool

« go back to overview

Description

general schooling history

File structure

spell format: 1 row = 1 school episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

73 / 16,734

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
subspell	Number of subspell
spell	Spell number
wave	Wave
ts11204	Type of school
ts1111m	Start date School episode, month
ts1111y	Start date School episode, year
ts1112m	End date School episode (month)
ts1112y	End date School episode (year)
ts11209	School-leaving qualification
ts11214	Intended school-leaving qualification

Exemplary data snapshot

ID_t	splink	subspell	spell	wave	ts1111y	ts1112y
4005942	220003	2	3	9	2015	2017
4007346	220003	2	3	9	2016	2018
4008142	220003	1	3	8	2016	2016
4040398	220004	3	4	10	2016	2018
4040672	220003	2	3	9	2016	2017

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.

Stata 27: Working with spSchool (find R example here)

```
** open the data file
use ${datapath}/SC3_spSchool_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.28 spSchoolExtExam

« go back to overview

Description

school exam certificates acquired outside of the regular German educational system

File structure

entity format: 1 row = 1 exam of 1 respondent

ID variables needed to identify a single row

ID_t exam

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

31 / 600

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
exam	Exam number
ts11300	Awarded qualification in Germany?
ts1130m	Date/month qualification was awarded
ts1130y	Date/year qualification was awarded
ts11302	Awarded school-leaving qualification
ts11300_g1	Awarded qualification in Germany? (edited)
ts11301_g1R	Country of awarded school-leaving qualification

Exemplary data snapshot

ID_t	wave	exam	ts11300	ts1130y	ts11302	ts11300_g1
4006212	10	1	1	2016	3	1
4007433	10	2	1	2017	7	1
4008246	10	1	1	2017	4	1
4010756	8	1	1	2015	7	1
4039812	10	1	1	2016	3	1

The file spSchoolExtExam comprises information about school exam certifications that have not been acquired through “regular” schooling in the German educational system. These can consist of:

- certificates that have been acquired abroad and were accredited by German authorities
- certificates that have been acquired in a German school as external examinee (i. e., without attending class lessons)
- certificates that are automatically awarded by advancing through grades in upper secondary education

Stata 28: Working with spSchoolExtExam (find R example here)

```
** aim of this example is to evaluate the age of the respondent
** at the exam

** first, we have to get the birth date of the respondent
use ID_t tx8050m tx8050y using ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** remove missing or irregular duplicates, so file becomes cross-sectional
nepsmiss tx8050m tx8050y
drop if missing(tx8050m) | missing(tx8050y)
duplicates drop ID_t, force

label language en
tempfile temp
save `temp'

** now, open the data file
use ${datapath}/SC3_spSchoolExtExam_D_${version}.dta, clear
label language en

** merge the previously extracted birth dates
merge m:1 ID_t using `temp', keep(master match) nogenerate

** recode the two date variables (year, month) into one:
gen exam_date=ym(ts1130y,ts1130m)
gen birth_date=ym(tx8050y,tx8050m)
format *_date %tm

** calculate the age (in years)
gen age=(exam_date-birth_date)/12

** recode missings to .a, b.,... (not necessarily needed)
nepsmiss ts11302

** show some deviation
tabulate ts11302, summarize(age)
```

4.5.29 spSibling

« go back to overview

Description

siblings of respondent

File structure

entity format: 1 row = 1 sibling of 1 respondent

ID variables needed to identify a single row

ID_t sibling wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

36 / 8,620

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
sibling	Sibling number
p732107	Sibling lives with parents
p73221m	Month of birth Sibling
p73221y	Year of birth Sibling
p732220	Gender Sibling
p732230	Relationship link sibling
p732313	Highest school-leaving qualification Sibling
p732314	Current vocationla training Sibling
p732315	Current civil service training Sibling
p732316	Type of attended higher education institution Sibling
p732324	Doctorate Sibling
p732325	Type of civil service training Sibling
p732401	Employment status Sibling
p732402	Unemployment Sibling

Exemplary data snapshot

ID_t	wave	sibling	p73221y	p732220	p732230
4007405	2	2	1994	1	biological brother/biological sister
4009109	2	2	1997	1	biological brother/biological sister
4010468	2	1	2004	1	biological brother/biological sister
4039866	3	1	2001	1	biological brother/biological sister
4040282	3	1	2001	1	biological brother/biological sister

The file `spSibling` contains all siblings of the respondent, **reported by the parent**. Each sibling is stored in one row, containing information about the date of birth (`p73221m/y`), gender (`p732220`),employment status (`p732401`), and highest degree (`p732313`).

Stata 29: Working with spSibling (find R example here)

```
** aim of this example is to evaluate the number of older and younger
** siblings of a respondent

** first, we have to get the birth date of the respondent
use ID_t tx8050m tx8050y using ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** remove missing or irregular duplicates, so file becomes cross-sectional
nepsmiss tx8050m tx8050y
drop if missing(tx8050m) | missing(tx8050y)
duplicates drop ID_t, force

label language en
tempfile temp
save `temp'

** now, open the spSibling data file
use ${datapath}/SC3_spSibling_D_${version}.dta, clear
label language en

** merge the previously extracted birth dates
merge m:1 ID_t using `temp', keep(master match) nogenerate

** recode the two date variables (year, month) into one:
gen sibling_bdate=ym(p73221y,p73221m)
gen target_bdate=ym(tx8050y,tx8050m)
format *_bdate %tm

** check the difference between the two
gen older=.
replace older=0 if sibling_bdate>target_bdate
replace older=1 if sibling_bdate<target_bdate
replace older=. if missing(sibling_bdate) | missing(target_bdate)

** care about twins. As we do not know the day (or even the hour),
** we can not know which is older. We set this for a missing thus.
replace older=. if (sibling_bdate==target_bdate)

** generate the total amount of older siblings
bysort ID_t: egen total_older=total(older)
** generate the total amount of younger siblings
bysort ID_t: egen total_younger=total(1-older)

** aggregate to a single line for each respondent.
** the file then is cross-sectional with ID_t the sole identificator
keep ID_t total*
duplicates drop
```

4.5.30 spUnemp

« go back to overview

Description

spell data on unemployment episodes

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

32 / 3,127

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
subspell	Number of subspell
spell	Spell number
wave	Wave
ts2511m	Start of period of unemployment (month)
ts2511y	Start of period of unemployment (year)
ts2512m	End Unemployment episode (month)
ts2512y	End Unemployment episode (year)
ts25202	Receipt unempl. benefits (stage I) or unempl. assistance from the start
ts25203	Registered unemployment currently the case/finished
ts25205	Number Job applications
ts25206	Invitations to job interviews
ts25207	Number job interviews

Exemplary data snapshot

ID_t	subspell	spell	wave	ts2511m	ts2511y	ts2512m	ts2512y
4006458	1	1	10	6	2018	10	2018
4007648	1	1	8	9	2016	1	2017
4007654	1	2	11	9	2019	10	2019
4008434	2	1	11	9	2018	11	2018
4027882	1	2	11	8	2019	1	2020

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 to both the beginning and the end of an unemployment spell.

Stata 30: Working with spUnemp (find R example here)

```
** open the data file
use ${datapath}/SC3_spUnemp_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.31 spVocExtExam

« go back to overview

Description

vocational education certificates acquired outside of the regular German educational system

File structure

entity format: 1 row = 1 exam of 1 respondent

ID variables needed to identify a single row

ID_t exam

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

21 / 14

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
exam	Exam number
ts15301_g1	Professional/specialization title (KldB 1988)
ts15301_g4	Professional/specialization title (ISCO-08)
ts15301_g6	Professional/specialization title (SIOPS-88)
ts1530m	Date (month) external examination
ts1530y	Date (month) external examination
ts15304	External examination qualification
ts15302	External examination in Germany/abroad

Exemplary data snapshot

ID_t	wave	exam	ts1530m	ts1530y	ts15304
4006926	11	1	7	2019	1
4008567	9	2	9	2017	28
4039653	11	1	3	2019	1
4040254	12	1	5	2020	28
4040981	12	1	8	2019	2

The file spVocExtExam comprises information about vocational training certifications that have not been received by “regularly” passing through the German vocational training system. These can consist of:

- certificates that have been acquired abroad and were accredited by German authorities
- certificates that have been acquired in a German vocational trainig exam as external examinee (i. e., without attending lessons or courses registered with German authorities)

This especially includes second and third state examinations for alumni of medicine and law studies.

Stata 31: Working with spVocExtExam (find R example here)

```
** aim of this example is to evaluate the age of the respondent
** at the exam

** first, we have to get the birth date of the respondent
use ID_t tx8050m tx8050y using ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** remove missing or irregular duplicates, so file becomes cross-sectional
nepsmiss tx8050m tx8050y
drop if missing(tx8050m) | missing(tx8050y)
duplicates drop ID_t, force

label language en
tempfile temp
save `temp'


** now, open the data file
use ${datapath}/SC3_spVocExtExam_D_${version}.dta, clear
label language en

** merge the previously extracted birth dates
merge m:1 ID_t using `temp', keep(master match) nogenerate

** recode the two date variables (year, month) into one:
gen exam_date=ym(ts1530y,ts1530m)
gen birth_date=ym(tx8050y,tx8050m)
format *_date %tm

** calculate the age (in years)
gen age=(exam_date-birth_date)/12

** recode missings to .a, b.,... (not necessarily needed)
nepsmiss ts15304

** show some deviation
tabulate ts15304, summarize(age)
```

4.5.32 spVocPrep

« go back to overview

Description

vocational preparation schemes

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

78 / 1,219

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
splink	Link for spell merging
spell	Spell number
subspell	Number of subspell
spgen	Generated spell
wave	Wave
ts13103	Type of measure
ts1311m	Start Vocational preparation - month
ts1311y	Start Vocational preparation - year
ts1312m	End Vocational preparation - month
ts1312y	End Vocational preparation - year
ts1312c	Continuation of the vocational preparatory year
ts13201	Termination vocational preparation

Exemplary data snapshot

ID_t	spell	subspell	wave	ts1311m	ts1311y	ts1312m	ts1312y
4005994	1	1	9	10	2017	11	2017
4006344	1	2	10	8	2017	6	2018
4007586	1	2	9	27	2016	27	2017
4008211	1	5	12	8	2016	6	2020
4010610	2	2	11	10	2017	7	2018

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.

Stata 32: Working with spVocPrep (find R example here)

```
** open the data file
use ${datapath}/SC3_spVocPrep_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.33 spVocTrain

« go back to overview

Description

vocational education history

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

198 / 12,977

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
spell	Spell number
subspell	Number of subspell
ts15201	Type of vocational training
ts1511m	Start date Vocational training episode (month)
ts1511y	Start date Vocational training episode (year)
ts1512m	End date Vocational training episode (month)
ts1512y	End date Vocational training episode (year)
ts15215	Company size of the training company
ts15219	Completion of vocational qualification
ts15221	Intended vocational qualification

Exemplary data snapshot

ID_t	spell	subspell	ts1511m	ts1511y	ts1512m	ts1512y
4010336	1	4	9	2016	10	2019
4009130	1	2	9	2019	11	2020
4006379	1	1	8	2019	11	2019
4008064	1	1	8	2018	12	2018
4006700	1	3	10	2017	11	2019

This module covers all further trainings, 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, or
  - 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` module 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 those modules. Interruptions of vocational training spells, so-called vocational interruption episodes, are stored in wide format (be aware of this when working with harmonized spell data!).

Stata 33: Working with spVocTrain (find R example here)

```
** open the data file
use ${datapath}/SC3_spVocTrain_D_${version}.dta, clear

** only keep full or harmonized episodes
keep if subspell==0

** save this file temporarily
tempfile tmp
save `tmp'

** open the Biography data file
use ${datapath}/SC3_Biography_D_${version}.dta, clear

** merge the previously created temporary data file to this
merge 1:1 ID_t splink using `tmp' , keep(master match)

** you now have an enhanced version of Biography, enriched by
** information from the spell module. The number of total episodes
** (i.e., the amount of rows in the Biography file) did not change.
** Verify this by tabulating the spell type by the merging variable
** generated during the merge process.
tab sptype _merge
```

4.5.34 TargetMethods

« go back to overview

Description

Paradata from the targets CATI/CAPI interview

File structure

long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_int

Number of variables / number of rows in file

24 / 26,962

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
cohort	NEPS Starting Cohort
tx80201	Interview: Survey mode (start)
tx80202	Interview: Survey mode (realized case)
tx80221	Interview: evaluable data set?
tx80301	Interviewer: gender
tx80303	Interviewer: highest school-leaving qualification
ID_int	Interviewer: ID
tx80205	Interview: interview interrupted
tx80200	Interview: number of all contact attempts

Exemplary data snapshot

ID_t	wave	tx80201	tx80221	tx80301	tx80303	ID_int
4007394	10	CATI	does apply	2	6	2826
4007748	10	CATI	does apply	1	17	2855
4007812	11	CATI	does apply	1	7	1062
4010903	11	CATI	does apply	2	7	2972
4040524	10	CATI	does apply	2	17	2828

This dataset offers a variety of information on the data collection during the CATI/CAPI interview, e. g., gender (tx80301) and age (tx80302) of the interviewer; interview duration (tx80209); response code (tx80207); and survey mode (tx80202).

Importantly, this file contains all contacted respondents whether an interview was realized or not (see variable tx80207 for more details). Thus, TargetMethods includes more cases than the data file pTarget.

Stata 34: Working with TargetMethods (find R example here)

```
** open the data file
use ${datapath}/SC3_TargetMethods_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** check out response code by wave
tab wave tx80207

** how many different interviewers did CATI surveys?
distinct ID_int

** get an overview on the count of contact attempts
summarize tx80200
```

4.5.35 Weights

« go back to overview

Description

Sample weights for various occasions

File structure

wide format: 1 row = 1 target

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

ID_i

Number of variables / number of rows in file

62 / 8,317

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
ID_i	Institution ID
sample	Sample that target person belongs to
stratum_exp	Explicit stratum of drawing sampling procedure
stratum_imp1	implicit stratum 1 (aggregated type of school of institution)
w_i	Design weight for institution
w_t	Design weight for target
w_t1	Cross-sectional weight for targets participating in wave 1
w_t2	Cross-sectional weight for targets participating in wave 2
w_t3	Cross-sectional weight for targets participating in wave 3

Exemplary data snapshot

ID_t	ID_i	sample	w_i	w_t	w_t1	w_t2
4009691	1000383	Main sample	94.79454	194.06329	1.17955	1.03604
4007763	1000544	Main sample	95.22131	208.44438	1.07541	0.90123
4008954	1000554	Main sample	92.04327	139.58307	0.47639	0.43948
4008557	1000570	Main sample	94.81999	138.21658	0.57956	0.48006
4009033	1000364	Main sample	119.02664	189.76699	1.03359	0.87062

Weighting variables (starting with `w_`) are included in the Weights dataset. Also, you find stratification (`stratum`) identifiers here. Given the quite complex structure of the sample, no final recommendations are at hand concerning the use of design and adjusted weights. More information about weight estimation can be found in the survey paper for waves 1 to 6 (Steinhauer and Zinn, 2016b) and the supplementing paper for waves 7 to 9 (Steinhauer and Zinn, 2017). There are no general rules available on how the use of design or adjusted weights render any possible analysis more stable. Weights may possibly help to highlight important features of the analysis, or at least serve as a robustness check for the performed analysis.

Stata 35: Working with Weights (find R example here)

```
** open Weights datafile
use ${datapath}/SC3_Weights_D_${version}.dta, clear

label language en

** note that this file is cross-sectional, although the weights
** seem to contain panel logic
isid ID_t
d w_t*

** only keep design weight
keep ID_t w_t

** create a "panel" logic, i.e., clone each row
expand 12

** then create a wave variable
bysort ID_t: gen wave=_n

** remove entries who are not part of the wave variable
drop if inlist(wave,4,6)

** save as temporary file
tempfile weights
save `weights', replace

** open CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear

** and merge weight
merge 1:1 ID_t wave using `weights', nogen

** note that this weight is only available for those
** respondents participated in all waves
tab wave tx80220 if !(w_t==0 | missing(w_t))
```

4.5.36 xPlausibleValues

« go back to overview

Description

Plausible Values of competence data

File structure

wide format: 1 row = 1 respondent

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

wave_w*

Number of variables / number of rows in file

509 / 7,674

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2010)
wave_w3	Row contains data from wave 3 (2012)
wave_w5	Row contains data from wave 5 (2014)
wave_w9	Row contains data from wave 9 (2016/2017)
mag5_pv1	Math: cross-sectional plausible value 1
mag5_pv2	Math: cross-sectional plausible value 2
mag5_pv10	Math: cross-sectional plausible value 10
mag5_pv1u	Math: longitudinal plausible value 1
mag5_pv2u	Math: longitudinal plausible value 2
mag5_pv10u	Math: longitudinal plausible value 10

Exemplary data snapshot

ID_t	wave_w1	wave_w3	mag5_pv1	mag5_pv2	mag5_pv10	mag5_pv1u
4009616	1	1	0.41951	0.65399	0.05632	0.47124
4008556	1	1	0.56052	0.39717	0.20098	1.10208
4006406	1	1	1.77779	0.50646	0.63163	1.39574
4005993	1	1	0.12787	0.39717	0.63616	1.21046
4007287	1	1	0.95871	1.02016	0.93097	1.43594

Plausible Values (PV) are a way of describing the competencies of individuals at the group level. They allow (unbiased) estimates of effects at the population level that are adjusted for measurement errors. In contrast to point estimators such as Weighted Likelihood Estimates (WLE), the use of Plausible Values is suitable for more precise inferential statistical tests in correlation and mean value analyses.

Plausible Values are based on the individual answers in the competence tests and additional background characteristics (e. g. gender, age, socioeconomic status). For each person, the probability distribution of his or her competence is first determined and then several values

are randomly drawn from it (hence *Plausible Values*). Hypothesis tests for the specific question of interest are calculated for each of these values and combined into an overall result.

Please find more information on Plausible Values in the corresponding NEPS Survey Paper (Scharl et al., 2020) and on our website:

→ www.neps-data.de > Data Center > Overview and Assistance > Plausible Values

Stata 36: Working with xPlausibleValues (find R example here)

```
** open datafile.
use ${datapath}/${cohort}_xPlausibleValues_D_${version}.dta, clear
label language en

** as the 'x' in the filename indicates, this is a cross sectional file
** (no wave structure). You can verify this by asking if one row is
** solely identified by the respondents ID
isid ID_t

** note that competence testing has been conducted in multiple waves.
** An indicator marks if a row contains information for a specific wave.
tab1 wave_w*

** see more on how to work with this data in the Survey Paper mentioned above!
```

4.5.37 xTargetCompetencies

« go back to overview

Description

Test data of respondents

File structure

wide format: 1 row = 1 target

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

wave_w*

Number of variables / number of rows in file

2,147 / 7,681

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2010)
wave_w5	Row contains data from wave 5 (2014)
dgci2101_sc3g5_c	DGCF (concluding): set 1 item 1
dgg5_sc3a	DGCF (perceptual speed): sum
dgg5_sc3b	DGCF (reasoning): sum
mag5d041_c	Mathematical competence: Item 1
mag5_sc1	Mathematical competence: WLE
mag5_sc2	Mathematical competence: SE(WLE)
efg10_sc1	English: WLE
efg10_sc2	English: standard error of WLE

Exemplary data snapshot

ID_t	wave_w1	wave_w5	dgg5_sc3a	dgg5_sc3b	mag5_sc1	mag5_sc2
4008112	1	1	38	7	0.23	0.55
4007376	1	1	44	11	0.99	0.52
4010487	1	1	62	5	1.37	0.57
4009615	1	1	43	9	1.08	0.53
4009104	1	1	50	10	0.25	0.47

The file `xTargetCompetencies` contains data from competence assessments conducted in regular schools. Scored item variables as well as scale variables are available in a cross-sectional format. See table 2 for which competencies have been tested and when those testings have been conducted. You can also use variables `wave_w*` to select rows only containing data from a specific wave.

Stata 37: Working with xTargetCompetencies (find R example here)

```
** open datafile
use ${datapath}/SC3_xTargetCompetencies_D_${version}.dta, clear

** change language to english (defaults to german)
label language en

** as the 'x' in the filename indicates, this is a cross sectional file
** (no wave structure). You can verify this by asking if one row is
** solely identified by the respondents ID
isid ID_t

** note that competence testing has been conducted in multiple waves
** an indicator marks if a row contains information for a specific wave
tab1 wave_w*

** to work with competence data, you might want to merge it to CohortProfile.
** if you want to keep the panel logic (and not only add all competencies
** to every wave), you need a mergeable wave variable in xTargetCompetencies.
** in this example, we focus on math competencies, which have been tested in wave 1.
generate wave=1

** now, remove cases which did not took part in the testing
drop if wave_w1==0

** and reduce the dataset to the relevant variables
keep ID_t wave mag9_sc1 mag9_sc2

** save a temporary datafile
tempfile tmp
save `tmp'

** and merge this to CohortProfile
use ${datapath}/SC3_CohortProfile_D_${version}.dta, clear
merge 1:1 ID_t wave using `tmp', nogen
```

4.5.38 xTargetCORONA

« go back to overview

Description

Data collected in May 2020 regarding the impact of the corona pandemic on respondents life

File structure

wide format: 1 row = 1 target

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

185 / 1,031

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12

Exemplary variables

ID_t	ID target
wave	Wave
t514001	Satisfaction with life
tm00001	Impact: coronavirus infection - no
tm00007	Impact: quarantine - no
tm00013	Employment status before coronavirus pandemic
tm00015	Systemically important profession
tm00016	Change working time
tm00017	Change work place

Exemplary data snapshot

ID_t	wave	t514001	tm00013	tm00015
4006388	.	8	I was employed	yes
4008764	.	7	I was employed	no
4009027	.	7	I was employed	yes
4010511	.	9	I was employed	no
4027724	.	10 completely satisfied	I was employed	no

This data have been established to investigate the medium and long-term effects of the corona pandemic on skills development and educational pathways over the life course. The following questions are in particular:

- How do learning environments change and which potentials and risks become clear through the beginning digitalization of learning?
- Are there effects on upcoming educational decisions and are there medium and long-term effects on social educational inequality
- What are the effects on educational outcomes, such as income, but also non-monetary returns, e. g., health and labor market participation

Data is collected by means of a cross-cohort questionnaire program adapted to the current situation of the respective participants. In order to collect this data in a timely manner, the

first questions were administered via online survey in the NEPS Starting Cohorts 2 to 6 in May 2020. As this time span did not overlap with regular survey waves, data from this survey is marked with a missing wave (`wave==.`), and is contained in this data file. The corresponding questions have then been integrated in an additional module on the Corona pandemic, which is part of the regular main surveys in all starting cohorts afterwards. You find these data in the file `pTarget`.

Stata 38: Working with xTargetCORONA (find R example here)

```
** open the file
use ${datapath}/${cohort}_xTargetCORONA_D_${version}.dta, clear
label language en


** note that the wave is missing,
** as this reflects the pre-wave survey in may 2020
tab wave


** but rows can be uniquely identified by ID_t and wave
isid ID_t wave
```

5 Special Issues

5.1 Introduction and life course concept

Starting in 2010 in grade 5, Starting Cohort 3 is the first NEPS cohort that allows a comprehensive analysis of both the educational trajectory through lower and upper secondary education and the transition to vocational training or study as well as into the labor market (Fabian et al., 2019; Wagner et al., 2019; Ludwig-Mayerhofer et al., 2019).

Starting Cohort 3 data contain rich information on students' school trajectories, personality traits, competencies, and plans for their future from fifth grade onward (collected prospectively in PAPI mode). As long as the students were in schools where the original sampling took place (see section 2.2), parents (in CATI mode) as well as teachers and school principals (in PAPI mode) were also interviewed as context persons. Thus, a comprehensive picture of students' educational decisions and trajectories can be obtained prospectively (and additionally retrospectively through the interviews with parents).

Upon leaving general education schools, the former students have been interviewed in CATI/CAPI plus CAWI mode from then on. CATI and CAPI interviews allowed us to implement a modularized life course measurement, which makes for a key advantage of the data.

Modularized life course measurement means that longitudinal information on the respondent's biography is collected through customized self-reports within predefined domains of life. These domains cover different kinds of activities, such as general schooling history, vocational training, employment or further training. For each domain, episodes are collected one after another in the life course interview. Thus, full personal biographies are recorded domain by domain. We will often refer to these life-domain-specific parts of the life course interview as "life-course modules" throughout this section.

The modularized life course measurement is a remarkable improvement in the collection of life course data as it implements key insights from cognitive psychology and neuroscientific research into survey research. The approach benefits the empirical analysis because it leads to more accurate and complete life course data (Ruland et al., 2016). Interviewee burden is reduced by pre-structuring life courses by separating them into life domains and thereby giving interviewees (more) easily accessible stimuli that strengthen their mental recalling and reporting of biographical events. As Drasch and Matthes, 2013 show, the modularized measurement leads, among others, to higher data quality, e.g., more reported unemployment episodes. In contrast to less structured calendar measurements that essentially ask what happened first, what next, what next, what next, the modularized approach reduces the risk that respondents forget or omit episodes, for instance, overlapping, parallel, or unpleasant ones – thus, it reduces streamlining of life courses and underreporting of life events. Thereby, survey researchers not only get more complete life course data but also more precise information, especially with regard

to dates and durations of certain activities. This is a great deal for longitudinal data collection and analysis because it reduces the measurement error and the confounding bias. Having more precise life course information on an independent variable, for instance, leads to less biased estimates of that variable as data is less noisy. Having more precise life course information in a dependent variable (e.g., in sequence or event history analyses) decreases the variance of the error terms – a fact that likewise strengthens causal analysis.

However, the life-domain-centered (i.e., modularized) approach has its downside, too. Above all – as the flip side of gaining more complete and less streamlined biographies – it leads to the reporting of more overlapping events across life domains. A respondent might, for instance, report having an employment of 32 hours per week while attending full-time vocational training at the same time or being on parental leave while being unemployed. In a second step of the life course interview – processed by software in the computer-assisted interview – the domain-specific life histories are therefore compiled into a full, cross-domain life course. This merging step includes coherence checks across life domains. For instance, when a respondent reports full-time employment parallel with full-time schooling, he or she is asked to sort this overlap out. In order to keep the individual checking of coherence in the life course data manageable and time-efficient, this verification is only applied to the occupational (esp., employment and unemployment) and the educational (esp. vocational training and further training) modules, but not to the further life-course modules like those on children.

A special feature of Starting Cohort 3 (and also of Starting Cohort 4) are the so-called "state-specific modules". After leaving the general school system, transitions from school to vocational training or to academic education and to work can occur at different points in time. As described before, the individual activity episodes are surveyed via the modularized life course measurement.

In addition, further information is collected prospectively and retrospectively on the school-to-work-transitions on a cross-sectional basis as close as possible. Under certain conditions (e.g., a new ongoing vocational training is reported in the wave), state-specific modules are therefore switched.

The compilation of the target groups for the state-specific modules is based on the information given in the modularized life course measurement. Therefore, the state-specific modules are all switched in the survey after the complete recording of the life course.

How life course information is compiled and checked across life domains (modules) in the Starting Cohort 3 life course interview is described in Ruland et al., 2016. In addition to Ruland et al., 2016 and the conceptual overview given in Ludwig-Mayerhofer et al., 2019, we provide an in-depth information about the Starting Cohort 3 life course measurement in this chapter (see also section 4.4 for more general information about NEPS episode data). In particular, we provide detailed information about the various modules of the life course interview as well as about the state-specific modules with a particular focus on key definitions and changes across waves respectively (see section 5.3 and table 9, table 10 for an overview).

The following tables give an overview of names, numbers, and main content of the life-course modules (table 9) and state-specific modules (table 10) in Starting Cohort 3. Module names and

numbers refer to the structuring of the questionnaire in the programming template. Please note that besides in this chapter, module numbers are only to be found in the field version of the survey instruments and in the field reports from the survey institutes. No reference to these numbers is used in other documentation, such as the Scientific Use File instruments, the codebooks, the variable search, or the datasets. For more information about the available documentation, see section 1.2. It should also be noted that the Scientific Use File contains more spell datasets than those listed in the following table (see figure 20).

Table 9: List of life-course modules in Starting Cohort 3

Module	Number	SUF files	Main contents
General Schooling	22 AS	spSchool	The module records primary and secondary education episodes.
Pre-Vocational Training	23 BV	spVocPrep	The module records episodes of vocational preparation measures.
Vocational Training	24 AB	spVocTrain	The module records all vocational or academic educational episodes.
Military	25 WD	spMilitary	The module records all episodes of military, civilian and voluntary services.
Employment	26 ET	spEmp	The module records information on all employment episodes that respondents report on gainful employment, e.g., all activities leading to income.
Unemployment	27 AL	spUnemp	The module records all current and past periods during which participants were unemployed, regardless of any registration with the Federal Employment Agency.
Further Training Activities	35 KU	spCourses	The course module records further training activities in the life-course modules.

(...)

Table 9: (continued)

Module	Number	SUF files	Main contents
	31 WB	spFurtherEdu1	The further training module records all further training activities that were not reported in the previous modules.
Children	29 KI	spChild spChildCohab	The modules record information on respondents' children and …
Parental Leave	29 EZ	spParLeave	… parental leave episodes.
Gap	50 LU	spGap	The gap module covers all temporal gaps between the main life course activities and collects the activities carried out within these gap periods.

Table 10: List of state-specific modules in Starting Cohort 3

Module	SUF files	Main contents
State-specific modules for transition into training		
40ÜM 64bFÖSQS	pTarget	The module records detailed retrospective data of transition activities immediately after leaving general school system.
41SozKap	pTarget	The module records prospective data on social capital with regards to the transition to vocational training/study for those who did not start these activities immediately after leaving general school system.
State-specific modules during training		
40bRC	pTarget	The module records data on educational decisions during training/studies/pre-vocational measures.
Online_TaskModule	pTarget	The module records various task types that describe the company-based part of vocational training.

(…)

Table 10: (continued)

Module	SUF files	Main contents
61aÜAM	pTarget	The module records data on the quality of training.
40cABretro	pTarget	The module records retrospective information of content that is normally asked during ongoing training in the vocational training module, but not in the case of short training episodes that were never observed while they were ongoing.
State-specific modules for transition to the labor market		
41SozKaplab	pTarget	The module records different prospective data about social capital with regard to transition from vocational training into the labor market.
61bÜAM	pTarget	The module records various prospective data relevant to future educational and career decisions.
62IAM	pTarget	The module records various retrospective data relevant to educational and career decisions.

5.2 Specifics of the different survey modes

As mentioned before, respondents were interviewed using different interview modes depending on whether they were still in the school context or not. In the following, we will explain how the change from in-school to out-of-school interviews took place and what differences exist between the initial interviews after leaving general education and the subsequent panel interviews.

5.2.1 Transition from in-school to out-of-school interviewing

Target persons of Starting Cohort 3 successively switched to the out-of-school CATI/CAPI surveys after leaving the general school system (see figure 21). As of 2015, after the end of grade 9, the switch was expected for the first students. Since only a small number of school leavers were expected in 2015 (wave 7), just a short CATI interview took place that year to check the status. A short screening at the beginning of the interview clarified whether the target person had already left general education or not.

Target persons who continued to attend general education but at whose school the NEPS surveys weren't (any longer) conducted (due to a change of general education school or because the school no longer participated in the NEPS study), were individually tracked or followed-up. These students were subsequently asked to participate in an online survey (CAWI). The online survey corresponds to the survey content from the PAPI questionnaire, which was given to students who still attended a general education school at which NEPS surveys were conducted.

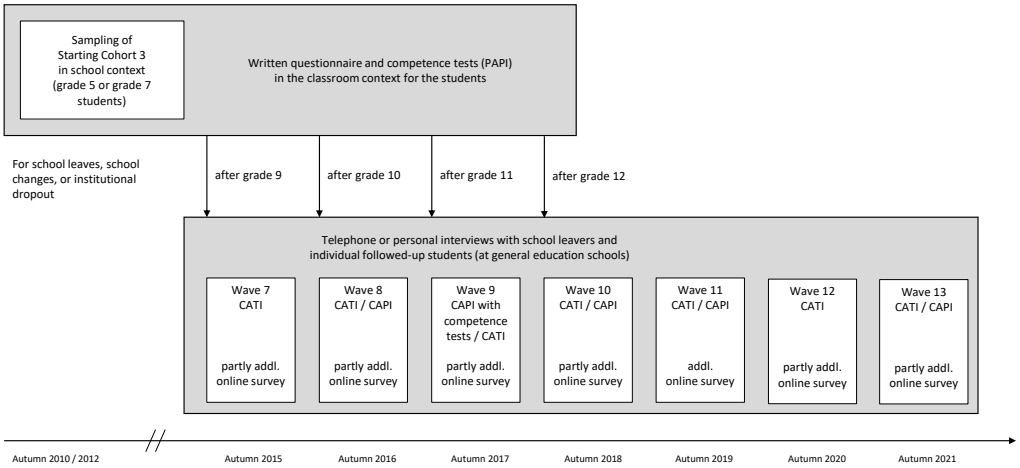

Figure 21: Transition from in-school to out-of-school-interviewing (adapted from Kersting and Aust, 2019)

Since 2016 (wave 8), school leavers had received a much more detailed interview. Furthermore, a short screening took place at the beginning of the interview. Individually tracked students received a shortened CATI interview and were afterwards again invited to an online survey to receive the question content that was simultaneously collected in the PAPI in the school context in the same grade. As of 2019 (wave 11), school surveys took no longer place and the entire Starting Cohort 3 was surveyed in an out-of-school context.

For the target persons who were interviewed for the first time in an out-of-school context, their school history and subsequent development after leaving school were surveyed retrospectively. For those who have already been interviewed repeatedly in an out-of-school context, this is a follow-up of their life course (see section 5.2.2).

There is also a follow-up online survey for certain groups. Students in upper secondary schools or at vocational grammar schools receive a survey instrument for students, adapted to their grade level, just like those who are individually tracked. Apprentices in their final year of training are asked about various tasks in their training. Since 2018 (wave 10), there has also been an online instrument for students. Variable `tx80230` in the `CohortProfile` dataset identifies if

the target person was interviewed as a student at a NEPS sample school (1), in an individual follow-up (2) or outside of the general school system (3).

5.2.2 Differences between initial survey and panel survey

In order to ensure that the retrospective record of the educational trajectory and the employment history is precise and complete, the survey is structured by life domains. The life history is split into different survey modules. Each of them covers the topic associated with that domain and captures corresponding activities, for example the (monthly) duration of school attendance.

In the initial survey, the entire biography of an interviewee is recorded retrospectively. Those initial surveys took place when respondents had left general education (see figure 21). In order to collect the biographical data, the activities within each module are recorded, starting with the first activity and ending with the current activities (the ongoing activities at the date of the interview, if applicable).

Once the biography was initially recorded, the respondent's biography is updated in each consecutive panel wave. Hence, the data from previous waves is used to adapt the questionnaire. Firstly, follow-up questions concerning activities recorded in the previous interview are asked. The interviewee can object to preloaded information in case it was recorded incorrectly in the earlier interview. Otherwise, the respective episode continues. Secondly, new activities are recorded that have started (and ended) since the last interview. Those new activities are also recorded chronologically per wave until the date of the current interview. Thereby, biography data are completed wave by wave, in each case referring to information from the previous wave.

5.3 Further information on data files

5.3.1 General schooling history

SUF file spVocTrain

Module 22 AS

This module captures the general schooling history from primary to upper secondary education.

The general schooling history module inquires about the type and location (municipality and federal state, as well as visits abroad) of the attended school(s); the nature of attendance (full-time, part-time); attained or aspired-to degrees; the accomplished grades of the degree and/or

during the last semester (in Math and German); so-called *Kopfnoten* (conduct and behavioral assessment) and missing hours (excused or unexcused) on the degree. For "Abitur" it asks for the examination subjects (up to five). For vocational schools it asks abouth practical education and whether it was in the form of an internship.

Special issues

Difference between retrospective and prospective schooling history

While respondents were still in school, the schooling history was collected via the parents. Once respondents have left school (or left a NEPS school), they were transferred into the individual CATI field where they were asked to retrospectively recall their entire schooling history. This happened with the aid of a looping questionnaire that chronologically records each schooling episode. For each schooling episode, the start and end dates and the type of school as well as its location were recorded. At the start of the module, respondents were also asked whether they had attended Kindergarten before primary school. Asking the respondents themselves had the advantage that a schooling history could now be collected for respondents whose parents did not participate in the earlier rounds or whose parents had provided incomplete information on the respondents schooling history.

An alternative way to generate a schooling history in the absence of parental information is to exploit information given by parents of the respondent's classmates (Bayer et al., 2014). However, this approach is not useful once respondent's leave the originally sampled school.

Important to note is that in the case of pupils switching school across federal states (*Bundesländer*), gaps in the schooling history might appear due to differences in the timing of summer vacations between the respective federal states.

5.3.2 Vocational preparation

SUF file spVocPrep

Module 23 BV

The vocational preparation module captures participation in vocational preparation measures (BV), including the *Berufsvorbereitungsjahr* (BVJ, vocational preparation year), the *Berufsgrundbildungsjahr* (BGJ, basic vocational education year), the one-year *Berufsfachschulausbildung*, or any other measure of vocational preparation by the employment agency such as *Einstiegsqualifizierung* (EQ/EQJ, entry qualification) or *berufsvorbereitende Bildungsmaßnahme* (BvB, vocationally preparing educational measure) or *Berufseinstiegsjahr* (BEJ, occupational entry year) or other measures that prepare for vocational training.

The module inquires about many aspects of vocational preparation measures (BV). Among others, there are detailed information on the type of BV; its location and duration; whether it is part-time or full-time; if the respondent has dropped-out of the BV and if so, the reasons for dropping out; which occupational field it is in and the reasons of choosing that field; whether it is possible to attain an educational degree and if so, the type of that degree; the place of the measure (school, company); whether it is in the form of an internship; and whether the respondent is assigned a case worker. Furthermore, the module captures subjective assessments of the learning progress during the BV; whether it has raised the respondent's interest; and whether the respondent feels like participation increased his/her employment chances. It ends with questions on the respondent's future plans after the BV.

5.3.3 Vocational training

SUF file spVocTrain

Module 24 AB

The vocational training module records data on vocational or academic education, including vocational training, college education or post-graduate degrees. If an individual participates in multiple educational activities at the same time, all activities are recorded as individual episodes. If an episode is no longer ongoing, the episode's end is the day of graduation or the day of dropout. Even if the educational activity is not completed, it is recorded in this module.

Amongst others, the vocational training module inquires about the type of vocational training taken; the location and duration of the training; the type of contract including salary; about various aspects on the funding of the training, the degree obtained in addition to the vocational training degree and the grade; about the satisfaction with the vocational training; about dropouts with very detailed information on determinants and reasons for dropping out of the educational activity.

For college degrees, a new episode is captured if the major subject or the college changes. Changes in minors or changes in the type of degree to be acquired are disregarded. For post-graduate degrees, characteristics of the degree are recorded.

In general, further training activities are recorded in the further training module or the course module (see section 5.3.7). In some cases, trainings are recorded in this vocational training module, for example IHK courses (*Industrie- und Handelskammer*, Chamber of Commerce and Industry).

After general vocational training, the module asks for further educational episodes in the context of external examinations (*Externenprüfungen*).

Special issues

Wave-specific

For wave 10, licensed courses and IHK courses are not captured in the vocational training module as it was before. The idea was that respondents report them directly in the further training module (see section 5.3.7) to save overall survey time. For wave 10, these courses are stored in the data file `spFurtherEdu1`. Unfortunately, it then turned out that this change has led to an underreporting of such courses. Thus, this process has been changed again for wave 11 and beyond to the following:

- IHK courses are again reported in the vocational training module and are hence stored in the `spVocTrain` dataset.
- Licensed courses are now captured in the course module (see section 5.3.7) and are therefore stored in the file `spCourses`.

5.3.4 Military

SUF file spMilitary

Module 25 WD

Contrary to the Starting Cohorts SC5 and SC6 (whose respondents are older), for Starting Cohort 3 this module only records voluntary services like voluntary social years, ecological years, European voluntary services, federal voluntary services, and voluntary military services or international youth voluntary services.

5.3.5 Employment

SUF files spEmp, pTarget

Modules 26 ET

In Starting Cohort 3 the longitudinal information on employment has two components: The information on employment episodes is covered by the annual employment module of the questionnaire (26 ET) and the data is stored in the `spEmp` data file. Additional information on absent

days due to illness (43cFehlET), job characteristics (65kJM), job tasks and skill mismatches is stored in `pTarget`.

Content in spEmp

The dataset `spEmp` captures information on all employment episodes that respondents reported on employment, e.g., all activities leading to income. This includes regular employment, but also traineeships, secondary jobs, as well as paid or unpaid internships after leaving school. Volunteering, military conscription, vocational training, vacation jobs and internships while still being at school or in vocational preparation schemes are not covered by the `spEmp` file. For these data see the other modules described in this section.

Due to the modular recording of the life course, the collected data on the employment situation is not restricted to one job but also contains data on jobs at the same time. There is no restriction to a specific number of parallel jobs.

In the questionnaire of the employment module, the introduction statements give specific anchors for respondents to report employments, including those that are parallel to other employments or to vocational training. There is an additional information for the interviewers that these employments can also include internships after leaving school. However, one has to note that these different employment types can be reported anytime in the course of the module. There is no specification that a regular or the main employment should be stated first. It is also important to know that the `spEmp` data file does not contain clear information from the respondents whether a job represents a main or a secondary employment or whether a job is a main or secondary activity in comparison to activities reported in the other life-course modules. Since such a decision highly depends on the research question of interest, we strongly recommend a conception-based and thorough edition of the employment episodes together with the life course data of Starting Cohort 3 that suits best the underlying research purpose. As a starting point, Rompczyk and Kleinert, 2017 provide an instruction on how to edit the life course data of Starting Cohort 6 which can be helpful for working with the life course data of Starting Cohort 3. More information on this in also provided in section 4.4 of this manual.

The employment episodes in `spEmp` cover information on the following topics:

- Occupation coded in different German and international classifications
- General employment type and detailed information about the employment type, e.g., supervision and management tasks, temporary employment, whether the reported employment is situated in the subsidized labor market, whether the reported job is contract work or seasonal employment
- Working hours, gross and net earnings for employees as well as the profit before and after taxes for self-employed persons
- Company characteristics and conditions for the participation in further training, courses and seminars

For target persons who are under the age of 21 and do not have completed a vocational training, there is a reduced version of the employment module. In the introduction statement, these respondents are specifically asked about internships, trainee programs, or vacation jobs. If they report any employment, they get a shorter version of the module with a focus on topics like their job title, start and end dates, working hours, place of work, and income. Those who get the shortened questionnaire can be identified in the `spEmp` dataset with the variable `tf23913`.

Changes over time in spEmp

Wave 13

`ts23228` "Master Professional" included for required educational qualifications

Wave 11

`ts23217` Seasonal work is no longer recorded as one continuous episode at a time, but each episode of seasonal work separately

Wave 10

New variables:

- `ts23208` Mini-jobs
- `ts23247` Termination of the job (dismissal/quitted)
- `ts23248` Chain contracts for fixed-term contracts

Other changes:

- `ts2333m` and `ts2333y` Items deleted in the questionnaire
- `ts23320`, `ts2332m`, `ts2332y`, `ts23244`, `ts23245`, `ts23246` Comprehensive filter adjustments in the questionnaire[16]
- `ts23552` No longer asked whether subsequent employment with the same employer was already reported
- `ts23229`, `ts23230`, `ts23231`, `ts23232`, `ts23233`, `ts23234`, `ts23243` Yearly updates introduced
- `ts23410--ts23546` In addition to the yearly income updates, for all completed episodes the full income information is collected at the end of the employment spell

Wave 9

`ts23215` The value 1 "ABM jobs [labor market measure jobs]" deleted in questionnaire
`tf23913` For respondents under the age of 21 and without a degree (`tf23913=1`), the introduction statements include probationary internships now

16 Complex filters in the questionnaire were simplified and made clearer, e.g., by filtering all respondents into a time stamp variable and then defining new groups for filtering into the next items. The adjustments also include corrections of filters for some of the mentioned variables.

Content in pTarget

A special feature of collecting information on employment in Starting Cohort 3 is the additional panel information that goes beyond the episode data in the spEmp data file and is stored in the pTarget dataset. However, this additional information is not available for every topic on employment in every wave. Table 11 shows all topics related to employment ever covered and for which waves the information is available. For example, in wave 13 (the last wave), questions on job characteristics, job tasks and skill mismatch were covered, whereas questions on time and performance pressure were covered only in the wave before. For absent days during employment, the target persons were asked in all waves to sum it up. The questions of the job tasks and the skill mismatch modules were posed as part of the online questionnaire (CAWI), the other modules as part of the telephone interview (CATI). It is important to note that this additional panel information is only collected for the main employment episode if there are several parallel employment episodes at the same time. It was up to the respondents to decide which episode this is.

While the modules on job tasks, skill mismatch and time and performance pressure were applied only once, the modules on absent days and job characteristics were included in the survey for several times. To allow for longitudinal analyses, the phrasing of the questions was held constant over time in these two modules. All instruments are provided as part of the data documentation (see section 1.2).

Table 11: Items on employment by wave

Wave	8	9	10	11	12	13
Annual questionnaire	x	x	x	x	x	x
Job Tasks						x
Skill Mismatch						x
Job characteristics			x	x	x	x
Time and performance pressure					x	
Absent days due to illness - employed	x	x	x	x	x	x

5.3.6 Unemployment

SUF files spUnemp, pTarget

Module 27 AL

The unemployment module records all episodes during which respondents were unemployed. Respondents are considered unemployed when they are registered as unemployed or when they are not working, but are actively seeking for work.

Content

When a respondent participates in the NEPS survey for the first time, the module records current and past unemployment episodes retrospectively over the entire life course. In subsequent waves, the module collects all current and past unemployment episodes since the last interview. In addition, the data provide further details on the unemployment episode, the application process and on further training during unemployment. The data file `spUnemp` contains information on the following topics:

- Registration of unemployment, receipt of unemployment benefit
- Number of job applications and job interviews
- Courses or further training during unemployment, financed by the employment agency
- Job search efforts in the last four weeks (file `pTarget`)
- Possibilities to start a new job within two weeks (file `pTarget`)

Special issues

The module does not distinguish between different types of unemployment. Hence, no new unemployment episode starts when changes occur (e.g., in registration of unemployment or benefit). Thus, there are no consecutive unemployment episodes as the module records the entire period of unemployment in one piece. Furthermore, the module does not record unemployment episodes for periods immediately before the end of training or employment, even if the respondent is already registered as a job seeker because of the three-month registration deadline in German employment agencies.

5.3.7 Further training activities

SUF files spCourses, spFurtherEdu1

Modules 35 KU, 31 WB

Further training activities are recorded throughout the questionnaire in two modules. First, the course module (35 KU) records further training in the life-course modules, such that respondents recall context-specific further training activities, for example courses taken during employment or during parental leave episodes. Second, the further training module (31 WB) captures all further training activities of the respondents in Starting Cohort 3, which have not been reported in earlier modules.

In addition, the vocational training module (see section 5.3.3) includes licensed courses and other vocational trainings which were identified as further training courses.

Content: Course module (spCourses)

When respondents state having participated in further training within an episode, this statement triggers the course module. For up to five courses per episode, the course module records the content, duration and completion. Since wave 10, there are additional questions on reasons (occupational vs. private), motivation, obligation and certification for the five courses. Further training activities are not recorded exactly to the date, but rather the respondents recall all training activities since the last interview or since the beginning of an episode.

Content: Further training module (spFurtherEdu1)

Towards the end of the interview, the further training module records all further training activities (classes, courses and seminars) in which the respondents participated since the last interview and that have not been reported yet. The module records all types of further training including courses taken out of personal interest, such as cooking or yoga classes.

At the beginning of the module, the interviewer reads all further training activities recorded in the course module to the respondent and asks whether the respondent has participated in any other further training activities since the last interview. The name of each further training activity is then captured along with its content, duration, and completion. Additionally, data on the reasons (private vs. occupational), the motivation, whether the class was mandatory, and whether the respondent received a certificate is recorded equally to the course module since wave 10.

Furthermore, additional data on randomly chosen courses are recorded within this module. For example, financial support for the course, the provider of the training, and respondent's evaluation of the course.

Irrespective of having participated in any further training activity, the module asks about participation in informal learning activities such as reading of specialized literature, using online learning tools or visiting trade fairs or specialized lectures. But in contrast to Starting Cohort 4 and above, there are no detailed take-up questions on this topic.

Special issues

Assignment of further training activities across waves

Further training activities that were not completed at the time of the interview are not incorporated in the next wave as a preload, which means that they might be reported again. It is not evident to assign a further training activity from the last wave to the next:

- The respondent would have to phrase the name of the further training activity exactly the same way in both waves.

- A respondent can report two different further training activities within the same field, even if the content and names of the further training activities are the same, for example two yoga classes.
- For the Scientific Use File in the download version: Only the categorical variable for further training course content (`tx272000_g13`) allows the assignment of a training activity from the previous wave to the next. However, due to data protection reasons, the course content is aggregated and categorized in this variable, therefore identically categorized courses have a high likelihood of actually being different further training activities.

5.3.8 Children and parental leave

SUF files Children, spChild, spChildCohab, spParLeave

Modules 29 KI, 29 EZ

Data on respondents' children and parental leave episodes are recorded throughout the questionnaire in two modules: First, the children module (29 KI) collects data on respondents' children and related living conditions. Second, the parental leave module (29 EZ) captures data on parental leave episodes as part of the life course.

Content: Children module

The children module is queried to all respondents of the study. However, respondents without children are only asked about their private or voluntary care services of, e.g., family members or friends. Respondents with children pass through the whole module. Here, data on all adopted, foster, and biological children are collected – including children living in the same household.

There is an item loop for every child. It consists of items on sociodemographic information, episodes of living together in one household, and episodes of parental leave.

If the respondent reports an episode of parental leave, a redirection to the parental leave module (see below) follows since wave 12, as well as a redirection to the course module (see section 5.3.7) in case of further training activities during this episode. Back in the children module, child-specific data on the childcare situation, respondent's educational aspirations for the child, the current activity status, and educational and vocational certificates are recorded depending on the child's age. At the end of the module, there are some general cross-sectional questions about the respondent's engagement in childcare and further care activities.

Content: Parental leave

The parental leave module was created in wave 12 as a decoupling of items from the children module. Respondents are redirected to the parental leave module if they indicate an episode of parental leave in the children module or in the data revision module. The episode dates are

recorded in the original module. Information on administrative issues, and the re-entry into employment are part of the parental leave module.

Since parental leave is recorded in the form of a life course episode, parental leave episodes are also considered during the life course check in the data revision module. Such an episode often runs parallel to other episodes, e.g., employment, even if the respondent was not working during the parental leave. This is due to the fact that in the interview all other life course episodes are collected before the recording of parental leave.

Changes over time

Wave 12

- The items on the career re-entry after parental leave are decoupled as a new module. This makes it possible to collect the information even if a parental leave episode is added to the life course later during the data revision or check module.
- The definition of parental leave has changed. Previously, the respondents were asked to indicate parental leave only if they had a legal right to this parental leave and did not work more than 30 hours per week. As of wave 12, the definition of a parental leave is up to the respondents.

Wave 11

- Items on the care situation of every child (as part of the child loop) are added to the module.

Wave 10

- For respondents who were not asked about children for the first time, questions regarding intentions of having children over the next two years were included in the children module.

5.3.9 Gap

SUF file spGap

Module 50 LU

Immediately after collecting the main life course activities of a respondent within the modules school, vocational preparation, vocational training, military, employment, unemployment, and parental leave, the data revision module checks, among other things, whether there are any chronological gaps between these main activities. If this is the case, these chronological gaps are closed in collaboration with the interviewee. This is done by either specifying an additional main activity by the interviewee that closes the gap, or by specifying an activity that is not covered by the main activities, a so-called “gap activity”. In the first case, the survey instrument

branches from the data revision module back to the corresponding module for the main activity. There, a new episode with a main activity will be collected to close the chronological gap in the life course. Then, the survey instrument filters back into the data revision module. In the second case, i.e., if no main activities were exercised in the chronological gap, the gap module becomes activated. Here, other activities can be specified to fill the gap, such as being *housewife/househusband*, *sick/unable to work*, etc. In this respect, unlike the main activity modules, the gap module is only used within the data revision module to fill in chronological gaps.

One exception to the closing of chronological gaps in the data revision module is the main activity "parental leave". Since the recording of parental leave episodes does not take place in a standalone module, but is embedded in the child module and is done there for each child separately, a direct return from the data revision module to the corresponding main activity module is not possible. Instead, an episode of parental leave recorded in the data revision module is treated as a gap activity. This means that instead of the main activity module, the gap activity module will be activated and the parental leave is recorded there, but without specific reference to a child and only regarding information on start and end date of the episode. As of survey wave 11, additional information on the employment of the interviewed person during this parental leave is collected in the gap module in the same way as in the parental leave module (see section 5.3.8).

Although the gap module is only used to fill gaps, it is possible that gap activities and main activities overlap chronologically. On the one hand, this can be the case if after collecting a gap activity, other activities are collected in the data revision module and these activities overlap chronologically confirmed by the interviewed person. On the other hand, there may be chronological overlaps if a gap activity persisted at the time of the interview and was pursued further in the subsequent survey wave. In this case, the gap episode is continued regardless of the existence of a gap for this period in the life course.

5.3.10 School-to-work transitions

After leaving the general schooling system, transitions from school to vocational training, academic education, or work can occur at different points in time. As described in the previous chapters, individual activity episodes such as general schooling, vocational training, and employment are surveyed via the modularized life course measurement. Data on the transitions between these episodes are collected in the so-called "state-specific modules". The construction of target groups for these state-specific modules is based on the information given about the activities in the modularized life course measurement.

In the following, the surveyed contents of the different school-to-work-transitions are presented in brief (for more information about the research approaches and background, see Ludwig-Mayerhofer et al., 2019).

a) Educational decision-making at the end of general schooling

Starting in grade 9 and as long as they were still students (and sometimes also beyond that), respondents were asked about their career orientation, application activities for training, and their aspirations for both their education and their career.

Table 12: Items about occupational orientation by wave and grade

Wave[a]	4	5	6	7	8	9	10	11[b]	12	13
Meaning of work (t66210a to t66210p)	x	x								
Idealistic career aspirations (tf00010, t31060a)	x	x	x	x	x	x		x		
Realistic career aspirations (tf00020, t31160a)	x	x	x	x	x	x		x		
Realistic professional aspirations (tf00200)		x		x	x	x				
Status of vocational training application (tf00040, tf00050, te11030, tf00030, tf0021a, tf0027a, tf0027b, tf0029a, tf0029b, tf0032a, tf0032b, tf00360)	x	x		x	x	x				
Plans for vocational training (apprenticeship) (tf00070, tf00260, tf0013a, tf0013b, tf0019a, tf0019b, tf0030a, tf0030b, tf0033a, tf0033b)		x	x	x		x				
Reasons for applying (tf0008a to tf0008j)		x								
Subjective probability of success of a vocational training position (tf00090, tf00100)		x		x	x	x				

(...)

Table 12: (continued)

Wave[a]	4	5	6	7	8	9	10	11[b]	12	13
Role models / vocational orientation about social environment (tf0011a to tf0011e, tf00160, tf0017a to tf0017e, tf00390, tf0040a to tf0040e)		x		x	x	x				
Subjective information about getting a training position (tf00120, tf00180)		x		x	x	x				
Information sources and search strategies (tf0006a to tf0006i, tf0023a to tf0023h, tf00240, tf00250)	x		x	x	x	x				
Future career plans (te11020, te11040)					x	x				
Application success (tf0028a, tf0028b, tf0031a, tf0031b, tf0034a, tf0034b, tf0035a to tf0035h, tf0021a to tf0021c)			x	x	x	x				

[a] If several variables are listed in one row, at least one of the variables of the construct was collected in the waves mentioned.

[b] In this wave, data was collected from the entire starting cohort.

b) Determinants of youth's placement within the VET system

SUF file pTarget

Module 40ÜM, 64bFÖSQS, 41SozKap

To collect rich data on the transition activities into training and studies (Vocational Education and Training, VET), various modules have been implemented.

In the first survey after leaving the general schooling system, the transition activities to vocational training and higher education were surveyed retrospectively in great detail. This included specific questions on various application activities for vocational training or university studies,

reasons for not applying, reasons for starting a vocational training or pre-vocational measure, willingness to relocate for studies or vocational training, and the kind of support that was available for the transition. In addition, there were questions on aspects of post-school transitions that are particularly relevant for former students with special educational needs. To ensure a good comparability within the starting cohort, this information was collected from the entire cohort.

The items `t291401` to `tf40192` and `tf15470` to `tf15488` can be found in the `pTarget` file. Within the modules, respondents are filtered according to their situation. The variables `tf40003` (entitled to study or currently studying: yes/no) and `tf40002` (current status: in vocational training, in transitional activity, other activity or studying) are crucial for filtering in the modules. All content is available at least once for all participating respondents from wave 8, with the beginning of the out-of-school interviews.

If no transition to vocational training or study had taken place in the follow-up waves either, respondents were asked retrospectively again about their transition activities.[17] This included questions about application activities for vocational training and university studies, reasons for not applying, willingness to relocate for studies or vocational training, and also what kind of personal support was available for the transition.

Those who did not transition to training and education immediately after leaving general education were prospectively surveyed about their social capital resources. This involved the likelihood of persons in the social environment of the interviewees providing information about training positions and offering help with applications, and if so, which ones.

The items `t32401*` and `t32501*` were collected from all respondents who had not started training or studying directly after leaving school, were not currently in training or studying at the time of the survey, and had not yet successfully completed training or studying.

Changes over time

Wave 10

`tf40101`, `tf40111` to `tf40124`: Items deleted in questionnaire

Wave 9

`tf40103` to `tf40107` (reasons for not applying), `tf40125`, `tf40126`: Filter adjustments in questionnaire[18]

17 To be more specific: The short version of questions about the transition activities into training and studies were collected by respondents who were neither in vocational training or studying nor currently attending upper secondary school, and haven't finished successfully a vocational training or study program at the time of the interview.
18 Complex filters in the questionnaire were simplified and made clearer.

c) (Un)Successful completion of VET programs

SUF file pTarget

Module 40bRC, 61aÜAM, 40cABretro, Online_TaskModul

There is little known about the determinants of the intra- and the inter-individual differences in the pathways through the VET system and their outcomes. To fill this knowledge gap, different modules have been implemented in the NEPS survey.

Whenever respondents reported having newly started a VET which is still ongoing at the time of the interview, they were asked about a large set of motivational factors and items on learning environments. Items `t30450a` to `t30740b` include data about how work climate, work composition, class climate and class composition are perceived at the vocational school. Also available is information on how important the respondent's occupational success is to parents and friends. Within the modules, respondents are filtered according to their situation. The variables `tf40001` (current status: in dual vocational education and training, in school-based training, or no training) and `tf40002` are crucial for filtering in these modules. Item `ts15106` in `spVocTrain` helps to understand the filtering with regard to the vocational training place (company-based, external or unknown, school-based or no training contract).

As long as the VET is ongoing, questions about training quality are also asked in each wave (items `t253001` to `t255054`). The questions cover various dimensions of training quality at the vocational school and in the training company.[19]

In order to examine the role played by the quality of training, but also by social capital resources in training in the case of an unsuccessfully completed training, a further module was implemented in Starting Cohort 3 in wave 10 (items `t321333` to `t321210` and `t321211` to `t321133`). In this module, the above-mentioned contents were collected for short training episodes that were reported for the first time at the time of the interview but had already been completed. In most cases, these are training dropouts at the beginning of vocational training.

A special feature of the NEPS data is the possibility of analyzing the correspondence between tasks learned and applied during training and in later occupations (Ludwig-Mayerhofer et al., 2019). The "job" task measurement in VET corresponds to the measurement of job tasks in later jobs (for information about the job task measurement see Matthes et al., 2014). In the final year of their training, respondents were asked to provide information on the various tasks they perform during their training (items `tf34300` to `tf34395`). This part of the survey always takes place after the CATI (or CAPI) interview in the CAWI mode.

19 In wave 9, these items were only collected from apprentices in their final year of apprenticeship. Since wave 10, the information has been collected from all persons with an ongoing vocational training.

d) Pathways from the VET system into the labor market

SUF file pTarget

Module 41SozKaplab, 61bÜAM, 62IAM

To gather detailed information on the transition from vocational training to work, different modules have been implemented. If respondents report to be in the last year of their vocational training, they were asked which persons from their personal environment would inform them about possible vacancies and who would support them in getting a new job in Germany (variables `t324030` to `t32503c`). Information was also collected on employment orientation, idealistic and realistic plans for the future, and future job search (variables `tf61107` to `tf61106`).

Respondents who indicated that they had successfully completed vocational training or a bachelor's or master's degree program for the first time since the last survey, were then asked again retrospectively about the above-mentioned content (variables `tf62100` to `tf62133`).[20] In addition to the questions about employment orientation and job search, reasons why the target person did not apply for jobs were also asked. This makes it possible to not only make detailed statements about the application behavior and concessions in the application process, but also to analyze what discourages people from applying for jobs. All data is collected in much greater detail retrospectively than prospectively.

20 That is, if a respondent has successfully completed vocational training and also finished a bachelor's degree program for the first time a few years later, he or she will be asked twice about these issues.

A References

Bayer, M., Goßmann, F., & Bela, D. (2014). *NEPS Technical Report: Generated school type variable t723080_g1 in Starting Chohorts 3 and 4* (NEPS Working Paper No. 46). Leibniz-Institut für Bildungsverläufe.

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition ZfE* (2nd ed.). Springer VS. https://doi.org/10.1007/978-3-658-23162-0

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

Drasch, K., & Matthes, B. (2013). Improving retrospective life course data by combining modularized self-reports and event history calendars: Experiences from a large scale survey. *Quality & Quantity*, *47 (2)*, 817–838. https://doi.org/10.1007/s11135-011-9568-0

Fabian, P., Goy, M., Jarsinski, S., Naujokat, K., Prosch, A., Strietholt, R., Blatt, I., & Bos, W. (2019). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 231–252). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0

FDZ-LIfBi. (2023). *Data Manual NEPS Starting Cohort 3–Grade 5, Paths Through Lower Secondary School, Scientific Use File Version 12.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hess, D., Steinwede, A., & Schneider, B. (2012). *Erhebung von retrospektiven Längsschnittdaten - Prüfmodul*. Bonn, infas Institut für angewandte Sozialwissenschaft GmbH.

Kersting, A., & Aust, F. (2019). *Methodenbericht. NEPS Startkohorte 3 (Schulabgänger und individuell nachverfolgte Schüler) – Haupterhebung Herbst 2018, Teilstudie B132*. Bonn, Germany: infas Institut für angewandte Sozialwissenschaft GmbH.

Künster, R. (2015a). *Startkohorte 6: Erwachsene (SC6) Datenversion 5.0.0. Technical Report 1: Edition und Korrektur der Lebensverlaufsdaten*. Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. Bamberg, Germany.

Künster, R. (2015b). *Startkohorte 6: Erwachsene (SC6) Datenversion 5.1.0. Technical Report: Korrektur der Lebensverlaufsdaten*. Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. Bamberg, Germany.

Ludwig-Mayerhofer, W., Pollak, R., Solga, H., Menze, L., Leuze, K., Edelstein, R., Künster, R., Ebralidze, E., Fehring, G., & Kühn, S. (2019). Vocational education and training and transitions into the labor market. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 277–295). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0

Matthes, B., Christoph, B., Janik, F., & Ruland, M. (2014). Collecting information on job tasks—an instrument to measure tasks required at the workplace in a multi-topic survey. *Journal for Labour Market Research*, *47*(4), 273–297. https://doi.org/10.1007/s12651-014-0155-4

Matthes, B., Reimer, M., & Künster, R. (2005). TrueTales – ein neues Instrument zur Erhebung von Längsschnittdaten. In *Arbeitsbericht 2 des Projektes „Frühe Karrieren und Familiengründung: Lebensverläufe der Geburtskohorte 1971 in Ost- und Westdeutschland“*.

Matthes, B., Reimer, M., & Künster, R. (2007). Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten. *Methoden, Daten, Analysen – Zeitschrift für Empirische Sozialforschung*, *1*(1), 69–92.

NEPS Network. (2022). *National Educational Panel Study, Scientific Use File of Starting Cohort Grade 5*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC3:12.0.0

NEPS Network. (2023). *Starting Cohort 3: Grade 5 (SC3), Wave 12, Questionnaires (SUF Version 12.0.0)*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). German National Educational Panel Study (NEPS). Bamberg.

Rompczyk, K., & Kleinert, C. (2017). *Episode-split biography data in NEPS starting cohort 6: structure and editing process* (NEPS Survey Paper 28). Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. Bamberg, Germany. https://doi.org/10.5157/NEPS:SP28:1.0

Ruland, M., Drasch, K., Künster, R., Matthes, B., & Steinwede, A. (2016). Data-Revision Module - A Beneficial Tool to Support Autobiographical Memory in Life-Course Studies. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study* (pp. 367–384). Springer VS.

Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6* (NEPS Survey Paper No. 10). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Scharl, A., & Zink, E. (2022). NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-scale Assessments in Education*, *10*(28). https://doi.org/10.1186/s40536-022-00145-5

Schönberger, K., & Koberg, T. (2017). *Regional Data: Microm*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Steinhauer, H. W., & Zinn, S. (2016a). *NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 3 of the National Educational Panel Study (Waves 1 to 5)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Steinhauer, H. W., & Zinn, S. (2016b). *NEPS Technical Report for Weighting: Weighting the sample of Starting Cohort 4 of the National Educational Panel Study (Wave 1 to 6)* (NEPS Survey Paper No. 2). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Steinhauer, H. W., & Zinn, S. (2017). *NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 4 of the National Educational Panel Study (Waves 7 to 9)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Wagner, W., Kropf, M., Kramer, J., Schilling, J., Berendes, K., Albrecht, R., Hübner, N., Rieger, S., Bachsleitner, A., Lühe, J., Nagy, G., Lüdtke, O., Jonkmann, K., Gruner, S., Maaz, K., & Trautwein, U. (2019). Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 253–276). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0

Wenzig, K. (2012). *NEPS-Daten mit DOIs referenzieren* (RatSWD Working Paper Series). Rat für Sozial- und Wirtschaftsdaten, Berlin.

Zielonka, M., & Pelz, S. (2015). *Implementation of the ISCED-97, CASMIN and Years of Education. Classification Schemes in SUF Starting Cohort 6*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

B Appendix

B.1 R examples

In this Appendix, you will find R usage examples that correspond to the Stata usage examples in the main body of the data manual. Just like there, the examples become more adaptable if some variables are defined beforehand:

```
# Starting Cohort
cohort  <- "3"

# version of this Scientific Use File
version <- "12-0-0"
```

To further ease the readability and shorten the examples, we also define a function `read.neps()`. Please note that you also need the libraries `readstata13` and (optionally) `Hmisc` for this to work. If you do not have this libraries installed on your computer, you can easily do so by using the command `install.packages("readstata13")`.

R 39: read.neps()

```
library(readstata13)
library(Hmisc)

## convenient wrapper function to 'read.dta13()'. Example of usage:
## cp <- read.neps("CohortProfile")
##
read.neps <- function(token,path="Z:/SUF/Download"){

  # absolute path to the file. Might need some adaption in your setting!
  # the current definition refers to
  # "Z:/SUF/Download/<cohort>/<cohort>_<version>/Stata14/
  #  <cohort>_<token>_<version>.dta"
  file <- paste0(
            path,"/",
            cohort,"/",
            cohort,"_",
            version,
            "/Stata14/",
            cohort,"_",
            token,"_",
            version,
            ".dta"
          )

  # read the data
  data <- read.dta13(file, convert.factors = F)

  # set the language to english (comment this out if you work in german)
  data <- suppressWarnings(set.lang(data, "en"))

  # The following step is not absolutely necessary.
  # However, it is recommended if you find it convenient to have the variable
  # labels handy during your analysis. After importing the dataset,
  # you can display an overview of all variable labels by running the command
  # 'varlabel(data)'. However, this command does not work anymore after modifying
  # the data, e.g., by deleting or merging variables, since the variable labels
  # are attached to the data frame, and not the single variable.
  # For this line to work, you need library(Hmisc) loaded.
  # Afterwards, you are able to show the label using the command 'label(..)'
  for(i in seq_along(data)){
    label(data[,i]) = attr(data,"var.labels")[i]
  }

  return(data)
}
```

R 40: Working with Biography

```
# import the data file
Biography <- read.neps("Biography")

# check out which spell modules you can merge to this file
addmargins(table(Biography$sptype))

# check that you will need splink to merge information
# from other modules to this file
anyDuplicated(Biography[,c("ID_t","splink")])
#returns "0" if there are no duplicates.
#If there are duplicates this command returns the index of the first duplicate
```

R 41: Working with EditionBackups

```
# In this example, we want to restore the original values in the variable
# t731353 (Highest professional qualification Father) of datafile pTarget

# oen the datafile
EditionBackups <- read.neps("EditionBackups")

# only keep rows containing data of the variable mentioned above
EditionBackups <- subset(EditionBackups,
                         EditionBackups$dataset == "pTarget" &
                         EditionBackups$varname == "t731353")

# check which variables we need for merging
table(EditionBackups$mergevars)


# then keep the merging variables  and the variable with
# the original values (for cross-checking, we also  keep the
# variable editvalue, which contains the values found in pTarget)
EditionBackups = subset(EditionBackups,
                        select = c(ID_t, wave, tx20100, sourcevalue_num, editvalue_
                          num))


# rename the variables to emphasize affiliation
names(EditionBackups)[names(EditionBackups) == "sourcevalue_num"] = "t731353_source"
names(EditionBackups)[names(EditionBackups) == "editvalue_num"] = "t731353_edit"

# open pTarget
pTarget <- read.neps("pTarget")


# add the data above
# After merging, Stata merge has one variable more than R, because in Stata
# a merge indicator is produced during the merging process and in R isn't.
# Since we need a merge indicator here, the merge command has to be extended:
pTarget = transform(merge(
  x = cbind(pTarget, source = "master"),
```

```
  #x contains the pTarget data set plus one extra column "source",
  #where source = "master"
  y = cbind(EditionBackups, source = "using"),
  #y contains the EditionBackups data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "wave","tx20100")),
  #merges x and y by ID_t and wave
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
                  #in the merged dataset, source = "both" if the observations is in x
                    AND in y
                  ifelse(!is.na(source.x), "master", "using")),
  #otherwise, source = "master" if the obs. is only in x
  #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

# check all editions made
View(subset(pTarget[c("ID_t", "wave", "t731353", "t731353_source", "t731353_edit")],
            pTarget$source == "both"))

# replace the variable in the datafile with its original value
pTarget[pTarget$source == "both",c("t731353")] <-
  pTarget[pTarget$source == "both",c("t731353_source")]
```

R 42: Working with Education

```
# we want to merge the school type from spSchool to this data file.
# For this to work, we first have to prepare spSchool and keep only
# harmonized episodes (subspell==0)
spSchool <- read.neps("spSchool")
spSchool <- subset(spSchool, subspell==0)

# now, open the Education data file
Education <- read.neps("Education")

# check out which spell modules you can merge to this file
table(Education$tx28100)

# only keep school episodes
Education <- subset(Education, tx28100==22)

# check that you will need splink to merge information
# from other modules to this file
anyDuplicated(Education[,c("ID_t","splink")])
# returns "0" if there are no duplicates.

# merge spSchool
Education <- merge(Education, spSchool, by= c("ID_t", "splink"), all=TRUE)
```

R 43: Working with ParentMethods

```
# open the data file
ParentMethods <- read.neps("ParentMethods")

# check out response code by wave
table(ParentMethods$wave, ParentMethods$px80207)

# how many different interviewers did CATI surveys?
length(unique(ParentMethods$ID_int))

# get an overview on the count of contact attempts
summary(ParentMethods$px80200)
```

R 44: Working with pCourseClass

```
# open the data file
pCourseClass <- read.neps("pCourseClass")

pCourseClass <- subset(pCourseClass,
                       # only keep recommended data rows
                       ex20100==1,
                       # reduce data file to some information
                       select = c(ID_cc, wave, e22940b, e227400_g1D)
                       )

# check uniqueness of key variables
anyDuplicated(pCourseClass[,c("ID_cc","wave")])
# returns "0" if there are no duplicates.

# open CohortProfile
CohortProfile <- read.neps("CohortProfile")

# add the class information to this file. Note that
# merge is m:1, as multiple students belong to the same class
CohortProfile <-
  merge(CohortProfile, pCourseClass, by= c("ID_cc", "wave"), all=TRUE)

# crosstab some variables
table(CohortProfile$t723080_g1, CohortProfile$e22940b)
```

R 45: Working with pCourseGerman

```
# open the data file
pCourseGerman <- read.neps("pCourseGerman")

pCourseGerman <- subset(pCourseGerman,
                       # only keep recommended data rows
                       ex20100==1,
                       # reduce data file to some information
                       select = c(ID_cg, wave, e538021)
                       )
```

```
# check uniqueness of key variables
anyDuplicated(pCourseGerman[,c("ID_cg","wave")])
# returns "0" if there are no duplicates.

# open CohortProfile
CohortProfile <- read.neps("CohortProfile")

# add the class information to this file. Note that
# merge is m:1, as multiple students belong to the same class
CohortProfile <-
  merge(CohortProfile, pCourseGerman, by= c("ID_cg", "wave"), all=TRUE)

# crosstab some variables
table(CohortProfile$t723080_g1, CohortProfile$e538021)
```

R 46: Working with pCourseMath

```
# open the data file
pCourseMath <- read.neps("pCourseMath")

pCourseMath <- subset(pCourseMath,
                      # only keep recommended data rows
                      ex20100==1,
                      # reduce data file to some information
                      select = c(ID_cm, wave, e538011)
                      )

# check uniqueness of key variables
anyDuplicated(pCourseMath[,c("ID_cm","wave")])
# returns "0" if there are no duplicates.

# open CohortProfile
CohortProfile <- read.neps("CohortProfile")

# add the class information to this file. Note that
# merge is m:1, as multiple students belong to the same class
CohortProfile <-
  merge(CohortProfile, pCourseMath, by= c("ID_cm", "wave"), all=TRUE)

# crosstab some variables
table(CohortProfile$t723080_g1, CohortProfile$e538011)
```

R 47: Working with pEducator

```
# Goal: Have class-teachers gender available in CohortProfile
# (i.e., on student level). Walkthrough:
# a) collect data from pEducator and simplify structure
# b) merge this data to pCourseClass, retaining easy structure
# c) merge combined data to CohortProfile

# a)
```

```
# open data from pEducator file
pEducator <- read.neps("pEducator")
pEducator <- subset(pEducator, select = c(ID_e, wave, e762110))

# this data file is still in panel logic (i.e., one row per wave), although
# the data itself is time-invariant! To ease later merging, we reduce
# complexity of this file by restructuring to a cross-sectional format.

# remove missing rows
pEducator <- subset(pEducator, e762110>0 & !is.na(e762110))

# remove duplicates; in case of discrepancy, keep data from first wave
pEducator <- pEducator[order(pEducator$ID_e,pEducator$wave),]
pEducator$wave <- NULL
pEducator <- pEducator[!duplicated(pEducator),]


# b)
# open class data file pCourseClass
pCourseClass <- read.neps("pCourseClass")
pCourseClass <- subset(pCourseClass,
                       # only keep recommended data rows
                       ex20100==1,
                       # reduce data file to some information
                       select = c(ID_cc, wave, ID_e)
)


# merge pEducator-data
pCourseClass <- merge(pCourseClass, pEducator, by= c("ID_e"), all=TRUE)


# c)
# open CohortProfile
CohortProfile <- read.neps("CohortProfile")

# merge the data
CohortProfile <- merge(CohortProfile, pCourseClass, by= c("ID_cc","wave"), all=TRUE)

# crosstab gender of child to gender of teacher
table(CohortProfile$tx80501, CohortProfile$e762110)
```

R 48: Working with pInstitution

```
# open the CohortProfile
CohortProfile <- read.neps("CohortProfile")

# open the institution-data
pInstitution <- read.neps("pInstitution")

# merge the size of the school to CohortProfile using school ID
size <- subset(pInstitution, select = c(ID_i, wave, h227100))
CohortProfile <- merge(CohortProfile, size, by= c("ID_i", "wave"), all=TRUE)
```

```
# recode all missing values (negative values) to NA
CohortProfile$h227100[CohortProfile$h227100<0] <- NA

# cluster the children according to the quantiles of the institution size
CohortProfile$quantile.size <- cut(CohortProfile$h227100, 5)

table(CohortProfile$quantile.size)
```

R 49: Working with pInstitutionMicrom

```
# open pTargetMicrom datafile. Note that this data file is only available OnSite!'
Microm <- read.neps("pInstitutionMicrom")

# additionally to ID_i and wave, line identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(Microm[,c("ID_i", "wave" ,"regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate


# tabulating wave against regio shows availability of all levels
# in wave 5 and 7, but only the most detailed level available
# in wave 1 and 3 (usually housing level)
addmargins(table(Microm$wave, Microm$regio))

# only keep housing level
Microm <- subset(Microm, Microm$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
Microm <- merge(CohortProfile, Microm, by = c("ID_i", "wave"), all = TRUE)
```

R 50: Working with pInstitutionRegioInfas

```
# open  datafile. Note that this data file is only available OnSite!
RegioInfas <- read.neps("pInstitutionRegioInfas")

# identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(RegioInfas[,c("ID_i", "regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate

# existing regional levels are:
table(RegioInfas$regio)

# only keep housing level
RegioInfas = subset(RegioInfas, RegioInfas$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
```

```
RegioInfas <- merge(CohortProfile, RegioInfas, by = c("ID_i","wave"), all = TRUE)
```

R 51: Working with pParent

```
## open the CohortProfile
CohortProfile <- read.neps("CohortProfile")

# and the pParent
pParent <- read.neps("pParent")
pParent <- subset(pParent, select = c(ID_t, wave, p731905, p731955))

# merge occupation of parents (both respondent and partner) from pParent
CohortProfile <- merge(CohortProfile, pParent, by= c("ID_t","wave"), all=TRUE)

# note that parent data is only available in certain waves
table(CohortProfile$p731905, CohortProfile$wave)

# thus, to work with this information in other waves, you
# first have to carry over the values to other rows
for (i in 2:length(CohortProfile$ID_t)) {
  if(CohortProfile$ID_t[i] == CohortProfile$ID_t[i-1]){
    if(is.na(CohortProfile$p731905[i])){
      CohortProfile$p731905[i] = CohortProfile$p731905[i-1]
    }
    if(is.na(CohortProfile$p731955[i])){
      CohortProfile$p731955[i] = CohortProfile$p731955[i-1]
    }
  }
}

# check the distribution of parents occupation in current type of school
table(CohortProfile$p731905, CohortProfile$t723080_g1)
table(CohortProfile$p731955, CohortProfile$t723080_g1)
```

R 52: Working with pTarget

```
## open the CohortProfile
CohortProfile <- read.neps("CohortProfile")

# as there are multiple instances of some IDs in a specific wave, we
# need this 'hack' to deduplicate the data during the following merge process
CohortProfile$tx20100 <- 1

# open pTarget
pTarget <- read.neps("pTarget")
pTarget <- subset(pTarget, select = c(ID_t, wave, t400500_g1, t400000_g1D))

# merge country of birth and generation status from pTarget
pTarget <- subset(pTarget, select = c(ID_t, wave, t400500_g1, t400000_g1D))
CohortProfile <- merge(CohortProfile, pTarget, by= c("ID_t","wave"), all=TRUE)
```

```
# recode missings
CohortProfile$t400500_g1[CohortProfile$t400500_g1<0] <- NA
CohortProfile$t400000_g1D[CohortProfile$t400000_g1D<0] <- NA

# note that parent data is only available in certain waves
table(CohortProfile$t400000_g1D, CohortProfile$wave)

# thus, to work with this information in other waves, you
# first have to carry over the values to other rows
for (i in 2:length(CohortProfile$ID_t)) {
  if(CohortProfile$ID_t[i] == CohortProfile$ID_t[i-1]){
    if(is.na(CohortProfile$t400000_g1D[i])){
      CohortProfile$t400000_g1D[i] = CohortProfile$t400000_g1D[i-1]
    }
    if(is.na(CohortProfile$t400500_g1[i])){
      CohortProfile$t400500_g1[i] = CohortProfile$t400500_g1[i-1]
    }
  }
}

# check the above alteration
table(CohortProfile$t400000_g1D, CohortProfile$wave)

# check the distribution between migration and current type of school
table(CohortProfile$t723080_g1, CohortProfile$t400000_g1D)
```

R 53: Working with pTargetMicrom

```
# open pTargetMicrom datafile. Note that this data file is only available OnSite!
Microm <- read.neps("pTargetMicrom")

# additionally to ID_t and wave, line identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(Microm[,c("ID_t", "wave" ,"regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate

# tabulating wave against regio shows availability of all levels
# in wave 5 and 7, but only the most detailed level available
# in wave 1 and 3 (usually housing level)
addmargins(table(Microm$wave, Microm$regio))

# only keep housing level
Microm <- subset(Microm, Microm$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
Microm <- merge(CohortProfile, Microm, by = c("ID_t", "wave"), all = TRUE)
```

R 54: Working with pTargetRegioInfas

```
# open RegioInfas datafile. Note that this data file is only available OnSite!
```

```
RegioInfas <- read.neps("pTargetRegioInfas")

# identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(RegioInfas[,c("ID_t", "regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate

# existing regional levels are:
table(RegioInfas$regio)

# only keep housing level
RegioInfas = subset(RegioInfas, RegioInfas$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
RegioInfas <- merge(CohortProfile, RegioInfas, by = c("ID_t","wave"), all = TRUE)
```

R 55: Working with spChild

```
# open the data file
spChild = read.neps("spChild")

# only keep full or harmonized episodes
spChild = subset(spChild, spChild$subspell ==  0)

# generate the total count of children for each respondent
# you can do this either by taking the maximum child number:
spChild = within(spChild, {children = ave(child, ID_t, FUN = max)})

# or counting the number of rows:
spChild = within(spChild, {children2 = ave(ID_t, ID_t, FUN = length)})

# which both computes the same result
identical(spChild$children, spChild$children2)

# recode rough values (e.g. end of year) to real months
spChild$ts3320m[spChild$ts3320m>20] <- spChild$ts3331m-20

# compute the age of one`s children today

#the zoo package is needed to transform time data
#install.packages("zoo")
library(zoo)

#transforms month names into month numbers
spChild$birth_ym = as.yearmon(paste(spChild$ts3320y, spChild$ts3320m), "%Y %m")

# then, create the same for the current date
spChild$today_ym = as.yearmon(rep(cut(Sys.Date(), "month"), length(spChild$ID_t)))

# the age is then easily computed
spChild$age = (spChild$today_ym - spChild$birth_ym)
```

```
summary(spChild$age)
# displays Min, Max and Mean of "age"
sd(spChild$age, na.rm = TRUE)
# displays Std.Dev. of "age"
length(spChild$age[!is.na(spChild$age)])
# displays the number of observations in "age" without NA
```

R 56: Working with spChildCohab

```
# open the data file
spChildCohab <- read.neps("spChildCohab")

# only keep full or harmonized episodes
spChildCohab <- subset(spChildCohab, spChildCohab$subspell ==  0)

# recode rough values (e.g. end of year) to real months
spChildCohab$ts3331m[spChildCohab$ts3331m>20] <- spChildCohab$ts3331m-20
spChildCohab$ts3332m[spChildCohab$ts3332m>20] <- spChildCohab$ts3332m-20


# generate the following durations in months:
# a) the total duration of a cohabitation episode

#the zoo package is needed to transform time data
#install.packages("zoo")
library(zoo)

spChildCohab$cohab_start =
  as.yearmon(paste(spChildCohab$ts3331y, spChildCohab$ts3331m), "%Y %m")
spChildCohab$cohab_end =
  as.yearmon(paste(spChildCohab$ts3332y, spChildCohab$ts3332m), "%Y %m")
spChildCohab$cohab_duration =
  (spChildCohab$cohab_end - spChildCohab$cohab_start)*12

# b) the total duration a respondent lived together with specific child
spChildCohab = within(spChildCohab,
                      {total_duration_per_child =
                        ave(cohab_duration, ID_t, child, FUN =
                              function(x) round(sum(x, na.rm = TRUE)))})


# c) the total duration a respondent lived together with any child
spChildCohab = within(spChildCohab,
                      {total_duration_per_target =
                        ave(cohab_duration, ID_t, FUN =
                              function(x) round(sum(x, na.rm = TRUE)))})

# to work with the latter information in other files, you could do
spChildCohab = subset(spChildCohab, select = c("ID_t", "total_duration_per_target"))
spChildCohab = spChildCohab[!duplicated(spChildCohab),]
# which gives you a cross-sectional display of cohabitation time per respondent
```

R 57: Working with spCourses

```
# open the data file
spCourses <- read.neps("spCourses")

# check which modules provided course information
cbind(addmargins(table(spCourses$sptype)))

# only keep courses from employment spells
spCourses <- subset(spCourses, spCourses$sptype ==  26)

# open the employment module
spEmp <- read.neps("spEmp")

# merge spCourses to spEmp
# note that this is an m:1 merge, as there are still subspells in spEmp
spEmp <-
  merge(spEmp, spCourses, by = c("ID_t", "wave", "splink"), all.x = TRUE)

# you now have the spEmp datafile, enhanced with information from spCourses,
# and can proceed with this in the usual way'
```

R 58: Working with spEmp

```
# open the data file
spEmp = read.neps("spEmp")

# only keep full or harmonized episodes
spEmp = subset(spEmp, spEmp$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge the spEmp to Biography
#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spEmp,source = "using"),
  #y contains the spEmp data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
    #in the merged dataset, source = "both" if the observations is in x AND in y
                  ifelse(!is.na(source.x), "master", "using")),
  #otherwise, source = "master" if the obs. is only in x
  #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
```

```
  #the columns "source" in x and y are deleted
)


#Since the variables wave and spms are in both data sets, Biography AND spEmp
#check before merging by: intersect(names(Biography), names(spEmp))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 59: Working with spFurtherEdu1

```
# open the datafile
spFurtherEdu1 <- read.neps("spFurtherEdu1")

# one row contains information for one course.
# The only possibility to use this file is to merge it to the data for this
# respondents wave (we use CohortProfile). So first, we have to remodel
# the file so one row contains one wave.
spFurtherEdu1$course_nr = ave(spFurtherEdu1$ID_t, spFurtherEdu1$ID_t,
                              spFurtherEdu1$wave, FUN = seq_along)

spFurtherEdu1 = reshape(data = spFurtherEdu1,
                        #data in long format
                        idvar = c("ID_t","wave"),
                        #idvar is/are the variable/s that need/s to be left unaltered
                        v.names = names(spFurtherEdu1[,3:16]),
                        #v.names contains names of variables in the long format that
                        #correspond to multiple variable in the wide format
                        timevar = "course_nr",
                        #timevar is/are the variable/s that need/s to be converted to
                        #wide format
                        direction = "wide")
#direction is to which format the data needs to be transformed


# open CohortProfile
CohortProfile <- read.neps("CohortProfile")


# merge the data
CohortProfile <-
  merge(CohortProfile, spFurtherEdu1, by = c("ID_t", "wave"), all.x = TRUE)
```

```
# Please note that you now have multiple variables added to CohortProfile,
# one set of variables for each course reported in spFurtherEdu1
```

R 60: Working with spGap

```
# open the data file'
spGap <- read.neps("spGap")

# only keep full or harmonized episodes
spGap = subset(spGap, spGap$subspell ==  0)

# open the Biography data file
Biography = read.neps("Biography")

# merge the spGap to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spGap,source = "using"),
  #y contains the spGap data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
               #otherwise, source = "master" if the obs. is only in x
               #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spGap
#check before merging by: intersect(names(Biography), names(spGap))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
```

```
addmargins(table(Biography$sptype, Biography$source))
```

R 61: Working with spMilitary

```
# open the data file
spMilitary <- read.neps("spMilitary")

# only keep full or harmonized episodes
spMilitary = subset(spMilitary, spMilitary$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spMilitary to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spMilitary,source = "using"),
  #y contains the spMilitary data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
           #otherwise, source = "master" if the obs. is only in x
           #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spMilitary
#check before merging by: intersect(names(Biography), names(spMilitary))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 62: Working with spParentGap

```
# open the data file
spParentGap <- read.neps("spParentGap")

# open the Biography data file
Biography <- read.neps("Biography")

# merge spParentGap to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spParentGap,source = "using"),
  #y contains the spParentGap data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
   #in the merged dataset, source = "both" if the observations is in x AND in y
                  ifelse(!is.na(source.x), "master", "using")),
  #otherwise, source = "master" if the obs. is only in x
  #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spParentGap
#check before merging by: intersect(names(Biography), names(spParentGap))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 63: Working with spParentSchool

```
# open the data file
spParentSchool <- read.neps("spParentSchool")
```

```
# only keep full or harmonized episodes
spParentSchool <- subset(spParentSchool, spParentSchool$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spParentSchool to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spParentSchool,source = "using"),
  #y contains the spParentSchool data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
    #in the merged dataset, source = "both" if the observations is in x AND in y
                  ifelse(!is.na(source.x), "master", "using")),
  #otherwise, source = "master" if the obs. is only in x
  #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND
  spParentSchool
#check before merging by: intersect(names(Biography), names(spParentSchool))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 64: Working with spParLeave

```
# open the data file
spParLeave <- read.neps("spParLeave")

# only keep full or harmonized episodes
```

```
spParLeave <- subset(spParLeave, spParLeave$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spParLeave to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spParLeave,source = "using"),
  #y contains the spParLeave data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
            ifelse(!is.na(source.x), "master", "using")),
            #otherwise, source = "master" if the obs. is only in x
            #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spParLeave
#check before merging by: intersect(names(Biography), names(spParLeave))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 65: Working with spSchool

```
# open the data file
spSchool <- read.neps("spSchool")

# only keep full or harmonized episodes
spSchool <- subset(spSchool, spSchool$subspell ==  0)
```

```
# open the Biography data file
Biography <- read.neps("Biography")

# merge spSchool to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spSchool,source = "using"),
  #y contains the spSchool data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
           #otherwise, source = "master" if the obs. is only in x
           #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spSchool
#check before merging by: intersect(names(Biography), names(spSchool))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 66: Working with spSchoolExtExam

```
# aim of this example is to evaluate the age of the respondent
# at the exam

# first, we have to get the birth date of the respondent
CohortProfile <- read.neps("CohortProfile")
CohortProfile <- subset(CohortProfile,
                        tx8050m>0 & !is.na(tx8050m) &
                        tx8050y>0 & !is.na(tx8050y),
```

```
                        select=c(ID_t,tx8050m, tx8050y))

# remove duplicates, so file becomes cross-sectional
CohortProfile <- CohortProfile[!duplicated(CohortProfile$ID_t),]

# now, open the data file
spSchoolExtExam <- read.neps("spSchoolExtExam")

# merge the previously extracted birth dates to spSchoolExtExam
spSchoolExtExam <-  merge(spSchoolExtExam, CohortProfile, by = c("ID_t"), all.x =
  TRUE)

# recode the two date variables (year, month) into one:
# the zoo package is needed to transform time data
#install.packages("zoo")
library(zoo)

spSchoolExtExam$exam_date =
        as.yearmon(paste(spSchoolExtExam$ts1130y, spSchoolExtExam$ts1130m), "%Y %m")
spSchoolExtExam$birth_date =
        as.yearmon(paste(spSchoolExtExam$tx8050y, spSchoolExtExam$tx8050m), "%Y %m")

# calculate the age (in years)
spSchoolExtExam$age = (spSchoolExtExam$exam_date - spSchoolExtExam$birth_date)


# show some deviation
aggregate(spSchoolExtExam$age, by = list(spSchoolExtExam$ts11302),
        FUN = function(x)
                c(mean = mean(x, na.rm = TRUE),
                sd = sd(x, na.rm = TRUE),freuquency = length(x[!is.na(x)])))
#displays mean and sd of age by school-leaving qualification

sum(!is.na(spSchoolExtExam$age))
#total number of observations without NA

summary(spSchoolExtExam$age)
#display mean of age in general

sd(spSchoolExtExam$age, na.rm = TRUE)
#display sd of age in general
```

R 67: Working with spSibling

```
# aim of this example is to evaluate the number of older and younger
# siblings of a respondent


# first, we have to get the birth date of the respondent
CohortProfile <- read.neps("CohortProfile")
CohortProfile <- subset(CohortProfile,
                        tx8050m>0 & !is.na(tx8050m) &
                          tx8050y>0 & !is.na(tx8050y),
```

```
                        select=c(ID_t,tx8050m, tx8050y))

# remove duplicates, so file becomes cross-sectional
CohortProfile <- CohortProfile[!duplicated(CohortProfile$ID_t),]


# now, open the spSibling data file
spSibling <- read.neps("spSibling")


# merge the previously extracted birth dates in pTargetCATI to spSibling
spSibling <- merge(spSibling, CohortProfile, by = c("ID_t"), all.x = TRUE)


# recode the two date variables (year, month) into one:
#the zoo package is needed to transform time data
#install.packages("zoo")
library(zoo)

spSibling$sibling_bdate =
        as.yearmon(paste(spSibling$p73221y, spSibling$p73221m), "%Y %m")
spSibling$target_bdate =
        as.yearmon(paste(spSibling$tx8050y, spSibling$tx8050m), "%Y %m")

# check the difference between the two

spSibling$older = rep(NA, times = length(spSibling$ID_t))
#create an empty variable "older"

#check the difference between the two bdates:
for (i in 1:length(spSibling$older)) {
  if(!is.na(spSibling$sibling_bdate[i]) & !is.na(spSibling$target_bdate[i]) &
            spSibling$sibling_bdate[i] > spSibling$target_bdate[i]) {
                spSibling$older[i] = 0
    } else {
      if (!is.na(spSibling$sibling_bdate[i]) & !is.na(spSibling$target_bdate[i]) &
                 spSibling$sibling_bdate[i] < spSibling$target_bdate[i]) {
                    spSibling$older[i] = 1
    } else {
      spSibling$older[i] = NA
    }
  }
}


# generate the total amount of older siblings
spSibling =
        within(spSibling, {total_older =
                ave(older, ID_t, FUN = function(x) sum(x, na.rm = TRUE))})

# generate the total amount of younger siblings
spSibling =
        within(spSibling, {total_younger =
                ave(older, ID_t, FUN = function(x) sum(1-x, na.rm = TRUE))})
```

```
# aggregate to a single line for each respondent.
# the file then is cross-sectional with ID_t the sole identificator
spSibling = subset(spSibling, select = c("ID_t", "total_older", "total_younger"))
#keep only the variables ID_t, total_older and total_younger

spSibling = unique(spSibling)
#drops duplicate rows from spSibling
```

R 68: Working with spUnemp

```
# open the data file
spUnemp <- read.neps("spUnemp")

# only keep full or harmonized episodes
spUnemp <- subset(spUnemp, spUnemp$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spUnemp to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spUnemp,source = "using"),
  #y contains the spUnemp data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
           #otherwise, source = "master" if the obs. is only in x
           #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spUnemp
#check before merging by: intersect(names(Biography), names(spUnemp))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL
```

```
# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 69: Working with spVocExtExam

```
# aim of this example is to evaluate the age of the respondent
# at the exam

# first, we have to get the birth date of the respondent
CohortProfile <- read.neps("CohortProfile")
CohortProfile <- subset(CohortProfile,
                        tx8050m>0 & !is.na(tx8050m) &
                          tx8050y>0 & !is.na(tx8050y),
                        select=c(ID_t,tx8050m, tx8050y))

# remove duplicates, so file becomes cross-sectional
CohortProfile <- CohortProfile[!duplicated(CohortProfile$ID_t),]


# open the data file spVocExtExam
spVocExtExam <- read.neps("spVocExtExam")

# merge the previously extracted birth dates in pTarget to spVocExtExam
spVocExtExam <- merge(spVocExtExam, CohortProfile, by = c("ID_t"), all.x = TRUE)


#the zoo package is needed to transform time data
#install.packages("zoo")
library(zoo)

# recode the two date variables (year, month) into one:
spVocExtExam$exam_date =
        as.yearmon(paste(spVocExtExam$ts1530y, spVocExtExam$ts1530m), "%Y %m")
spVocExtExam$birth_date =
        as.yearmon(paste(spVocExtExam$tx8050y, spVocExtExam$tx8050m), "%Y %m")

# calculate the age (in years)
spVocExtExam$age = (spVocExtExam$exam_date - spVocExtExam$birth_date)

# show some deviation
aggregate(spVocExtExam$age, by = list(spVocExtExam$ts15304),
        FUN = function(x)
                c(mean = mean(x, na.rm = TRUE),
                sd = sd(x, na.rm = TRUE),freuquency = length(x[!is.na(x)])))
#displays mean and sd of age by school-leaving qualification


sum(!is.na(spVocExtExam$age))
```

```
#total number of observations without NA

summary(spVocExtExam$age)
#displays mean of age in general

sd(spVocExtExam$age, na.rm = TRUE)
#displays sd of age in general
```

R 70: Working with spVocPrep

```
# open the data file
spVocPrep <- read.neps("spVocPrep")

# only keep full or harmonized episodes
spVocPrep <- subset(spVocPrep, spVocPrep$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spVocPrep to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spVocPrep,source = "using"),
  #y contains the spVocPrep data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
           #otherwise, source = "master" if the obs. is only in x
           #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spVocPrep
#check before merging by: intersect(names(Biography), names(spVocPrep))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
```

```
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 71: Working with spVocTrain

```
# open the data file
spVocTrain <- read.neps("spVocTrain")

# only keep full or harmonized episodes
spVocTrain <- subset(spVocTrain, spVocTrain$subspell ==  0)

# open the Biography data file
Biography <- read.neps("Biography")

# merge spVocTrain to Biography

#After merging, Stata merge has one variable more than R, because in Stata
#a merge indicator is produced during the merging process and in R isn't.
#Since we need a merge indicator here, the merge command has to be extended:
Biography = transform(merge(
  x = cbind(Biography,source = "master"),
  #x contains the Biography data set plus one extra column "source",
  #where source = "master"
  y = cbind(spVocTrain,source = "using"),
  #y contains the spVocTrain data set plus one extra column "source",
  #where source = "using"
  all.x = TRUE, by = c("ID_t", "splink")),
  #merges x and y by ID_t and splink
  source = ifelse(!is.na(source.x) & !is.na(source.y), "both",
  #in the merged dataset, source = "both" if the observations is in x AND in y
           ifelse(!is.na(source.x), "master", "using")),
           #otherwise, source = "master" if the obs. is only in x
           #and source = "using" if the obs. is only in y
  source.x = NULL,
  source.y = NULL
  #the columns "source" in x and y are deleted
)

#Since the variables wave and spms are in both data sets, Biography AND spVocTrain
#check before merging by: intersect(names(Biography), names(spVocTrain))
#and since the variables are not part of the merging process,
#both versions are contained in the new merged data set as
#wave.x/wave.y and spms.x/spms.y
#For each variable, one of the versions can be dropped by:
Biography$wave.y = NULL
Biography$spms.y = NULL

# you now have an enhanced version of Biography, enriched by
# information from the spell module. The number of total episodes
# (i.e. the amount of rows in the Biography file) did not change.
```

```
# Verify this by tabulating the spell type by the merging variable
# generated during the merge process.
addmargins(table(Biography$sptype, Biography$source))
```

R 72: Working with TargetMethods

```
# open the data file
TargetMethods <- read.neps("TargetMethods")

# check out response code by wave
table(TargetMethods$wave, TargetMethods$tx80207)

# how many different interviewers did CATI surveys?
length(unique(TargetMethods$ID_int))

# get an overview on the count of contact attempts
summary(TargetMethods$tx80200)
```

R 73: Working with Weights

```
# open Weights datafile
Weights <- read.neps("Weights")

# note that this file is cross-sectional, although the weights
# seem to contain panel logic
anyDuplicated(Weights[,c("ID_t")])
# returns "0" if there are no duplicates.

# only keep design weight
Weights <- subset(Weights, select = c(ID_t, w_t))

# open CohortProfile
CohortProfile <- read.neps("CohortProfile")

# and merge weight
CohortProfile <- merge(CohortProfile, Weights, by= c("ID_t"), all=TRUE)
```

R 74: Working with xPlausibleValues

```
# open datafile.
xPlausibleValues <- read.neps("xPlausibleValues")

# as the 'x' in the filename indicates, this is a cross sectional file
# (no wave structure). You can verify this by asking if one row is
# solely identified by the respondents ID
anyDuplicated(xPlausibleValues[,c("ID_t")])
# returns "0" if there are no duplicates.
# If there are duplicates this command returns the index of the first duplicate

# note that competence testing has been conducted in multiple waves.
```

```
# An indicator marks if a row contains information for a specific wave.
table(xPlausibleValues$wave_w1)

# see more on how to work with this data in the Survey Paper mentioned above!
```

R 75: Working with xTargetCompetencies

```
# open the data file
xTargetCompetencies <- read.neps("xTargetCompetencies")

#open the data file Cohort Profile
CohortProfile <- read.neps("CohortProfile")

# as the x in the filename indicates, this is a cross sectional file
# (no wave structure). You can verify this by asking if one row is
# solely identified by the respondents ID
anyDuplicated(xTargetCompetencies[,c("ID_t")])
# returns "0" if there are no duplicates.
# If there are duplicates this command returns the index of the first duplicate

# note that direct measures have been conducted in multiple waves.
# an indicator marks if a row contains information for a specific wave
table(xTargetCompetencies$wave_w1)
table(xTargetCompetencies$wave_w2)

# to work with competence data, you might want to merge it to CohortProfile.
# if you want to keep the panel logic (and not only add all competencies
# to every wave), you need a mergeable wave variable in xTargetCompetencies.
# in this example, we focus on math competencies, which have been tested in wave 1.
xTargetCompetencies$wave <- 1

# now, remove rows which do not hold relevant information
xTargetCompetencies <- subset(xTargetCompetencies, wave_w1 == 1)

# and reduce the dataset to the relevant variables
xTargetCompetencies <-
        subset(xTargetCompetencies, select = c(ID_t, wave, mag9_sc1, mag9_sc2))

# and merge the xDirectMeasures to CohortProfile
CohortProfile <-
        merge(CohortProfile, xTargetCompetencies, by= c("ID_t", "wave"), all=TRUE)
```

R 76: Working with xTargetCORONA

```
# open the file
xTargetCORONA <- read.neps("xTargetCORONA")

# note that the wave is missing,
# as this reflects the pre-wave survey in may 2020
table(xTargetCORONA$wave)

# but rows can be uniquely identified by ID_t and wave
```

```
anyDuplicated(xTargetCORONA[,c("ID_t","wave")])
# returns "0" if there are no duplicates.
```

B.2 Release notes

The following is the release note taken from the documentation page at the time this document has been computed:

```
================================================
**
** NEPS STARTING COHORT 3 - RELEASE NOTES a.k.a CHANGE LOG
** changes and updates for release NEPS SC3 12.0.0
** (doi:10.5157/NEPS:SC3:12.0.0)
**
================================================

* Known Issues *

CohortProfile:
        - variable t723080_g1 ('Current type of school (reconstructed)') for wave 6
            currently does not correctly distinguish
                between branches in schools with several educational programs; this
                    leads to the school type of students in
                such schools being erroneously sclassified as 'School with several
                    educational programs: Unclear' [value 8];
                this will be fixed in an upcoming release

xTargetCompetencies:
        - a few scientific literacy test items of wave 6 have incorrect value labels:
                => variables scg6103s_c, scg6142s_c, scg6144s_c, scg6664s_c, scg6111s_c
                     and scg6061s_c should be correctly
                        labeled 0 "not solved" and 1 "solved"
                        => in Stata you can easily relabel them by using the following
                             Stata syntax:
                            label values scg6103s_c  scg6142s_c scg6144s_c scg6664s_c
                                scg6111s_c scg6061s_c ':value label scg61050_c'
                => variable scg6661s_c should be correctly labeled with 0 "0 of 3
                     points", 1 "1 of 3 points", 2 "2 of 3 points",
                        and 3 "3 of 3 points"
                        => in Stata you can easily relabel it by using the following Stata
                             syntax:
                            label values scg6661s_c ':value label scg9012s_sc3g9_c'
                => variable scg6113s_c should be correctly labeled with 0 "0 of 2
                     points", 1 "1 of 2 points", 2 "2 of 2 points"
                        => in Stata you can easily relabel it by using the following Stata
                             syntax:
                            label values scg6113s_c ':value label scg11032s_sc3g11_c'

        - the linked WLEs for mathematical competence in Grade 9 (mag9_sc1u and
            mag9_sc2u) are currently incorrect due to a
                mistake in the linking procedure; this affects longitudinal comparisons
                     of means for Grade 9 which should
                currently not be conducted; the other linked WLEs for mathematical
                    competence in this cohort are not affected
                by this error; corrected WLEs in Grade 9 will be published with the
                    next SUF release


================================================
* Changes introduced to NEPS:SC3 by version 12.0.0 *
================================================
```

```
pTarget:
        - the wrong labels of the variables t10000a to t10000e for waves 7 to 11 from
            the last Scientific Use File releases
                have been corrected; the respective variables *_v1 have been deleted,
                    since the corresponding information
                for waves 1 to 6 is now included in the variables t10000a to t10000e

xTargetCompetencies:
        - the wrong variable name for mag9q021_c has been changed to mag9q171_c
            according to the naming conventions for
                competence variables

xTargetCORONA:
        - this dataset was renamed from pTargetCORONA to xTargetCORONA and contains
            only variables from the additional
                CAWI survey in May 2020 on the Corona pandemic; the Corona-specific
                    information from the questionnaire
                of the regular survey wave 12 were integrated into the dataset pTarget


================================================
* Changes introduced to NEPS:SC3 by version 11.0.1 *
================================================

spParentSchool:
        - the episode harmonization procedure, which combines information from multiple
             subspells into one harmonized spell
                (subspell==0), failed in the last version of the Scientific Use File
                    (11.0.0); this problem has been fixed
                with that update


================================================
* Changes introduced to NEPS:SC3 by version 11.0.0 *
================================================

General:
        - some variables were renamed for reasons of consistency with the Scientific
            Use Files of other NEPS starting cohorts
        - all spell datasets contained minor errors in the start and end dates of
            harmonized spells (subspell==0) for revoked
                episodes; this has been corrected

Biography:
        - episodes with missing information in both the start and the end date
            variables were excluded from the Biography dataset

CohortProfile:
        - the variable tx80230 ('panel frame') suffered from a coding error for the
            data of waves 6 to 10; this has been corrected

        - the former variables tx8600m/y, indicating the first interview date, were
            renamed to tx8601m/j ('Survey Target Person:
                survey month 1/year 1') for reasons of consistency with the Scientific
                    Use Files of other NEPS starting cohorts

pTarget:
        - the variables t520000 ('weight in kg') and t520001 ('height in cm') suffered
            from a coding error in the data of wave 10,
                which resulted in the values of the two variables being interchanged;
                    this has been corrected
```

```
TargetMethods:
        - variable tx80302 ('Interviewer: age group') suffered from a coding error;
            this has been corrected

xTargetCompetencies:
        - some variables including the item scores of the science test administered in
            grade 9 were erroneously excluded from the
                former SUF version 10.0.0; these variables are included again in the
                    current SUF version

        - an error has been corrected in the linkage scores for the tests of "
            Linguistic competence English: Reading"
                (efg10_sc1u and efg12_sc1u) which resulted in invalid longitudinal mean
                    -level comparisons across grades; the
                cross-sectional test scores were not affected


================================================
* Changes introduced to NEPS:SC3 by version 10.0.0 *
================================================

General:
        - all variables related to the date of data collection (i.e. when the
            competency tests and CATI interviews took place)
                have been updated and are now centrally stored in the CohortProfile
                    dataset; the variables intm and inty have
                been removed from all other datasets

pTarget:
        - the variables t66800l_g1, t66800m_g1, t66800n_g1, t66800q_g1 and t66800r_g1
            were newly generated to represent the
                measurement of the Big Five in wave 10 with 21 instead of 11 items; a
                    total of 10 items of the previous Big
                Five instrument from the waves 3 and 5 are part of the new measurement;
                     only the variable t66800k is not
                included in the new instrument, so that the original index variable
                    t66800b_g1 is not filled for wave 10

pTargetCorona:
        - a new dataset with information from an additional CAWI survey (May 2020) on
            Corona related topics has been
                incorporated in this SUF release

CohortProfile:
        - the missing information in wave 4 regarding the date of completion of the
            questionnaires by the students was added

xTargetCompetencies:
        - the competency scores for Reading Speed measures in wave 1 have been updated

spVocTrain:
        - in wave 9, questions were asked to collect information that was originally
            intended exclusively for Starting
                Cohort 4 and that is redundant and dispensable for Starting Cohort 3;
                    these variables [ts15550 - ts15559,
                ts15591_g1 - ts15591_g16] have been removed in this SUF release


================================================
* Changes introduced to NEPS:SC3 by version 9.0.0 *
================================================
```

```
General:
        - minor errors in variable labels have been corrected

        - the following datasets are prepared and published for the first time:
            Education and spVocExtExam; detailed information
                on these datasets can be found in the soon to be published data manual


================================================
* Changes introduced to NEPS:SC3 by version 8.0.1 *
================================================

General:
        - minor errors in variable labels have been corrected

pTarget:
        - the variables 'Gender' [t700031], 'Date of birth - month' [t70004m] and 'Date
             of birth - year' [t70004y] contained a
                coding error for wave 8 data; this error has been fixed

        - the generated variables 'Learning motivation school' [t66400a_g1 - t66409a_g1
            ], 'Learning motivation vocational
                training' [t66410a_g1 - t66413a_g1] and 'Learning motivation vocational
                     preparation' [t66415a_g1 - t66418a_g1]
                were newly added to the dataset


================================================
* Changes introduced to NEPS:SC3 by version 8.0.0 *
================================================

General:
        - the following datasets are prepared and published for the first time:
            Biography, EditionBackups, spChild, spChildCohab,
                spCourses, spEmp, spFurtherEdu1, spGap, spMilitary, spParLeave,
                    spSchool, spSchoolExtExam, spUnemp, spVocPrep,
                and spVocTrain; detailed descriptions of these datasets can be found in
                     the upcoming Data Manual

        - due to parental consent withdrawn between the survey waves, not all target
            persons' observations can usually be
                published in the Scientific Use File; in the last release (7.0.0,
                    7.0.1) data of target persons were inadvertently
                not included in the Scientific Use File, for which only the consent for
                     the parent survey was withdrawn, but not
                for the publication of the data by the target persons; the erroneously
                    dropped observations were now integrated
                again and the weights are re-calculated accordingly


================================================
* Changes introduced to NEPS:SC3 by version 7.0.1 *
================================================


Weights:
        - joint weights for students and parents have not been available in recent
            releases;
                they have been calculated by now and have been integrated into this
                    release

xTargetCompetencies:
```

```
        - uncorrected WLE score for mathematical competence in Grade 9 [mag9_sc1u] has
            been incorrectly linked; it should
                not been used for mean-level comparisons with preceding grades (
                    longitudinal correlational analyses or
                cross-sectional analyses are not affected); correctly linked WLE scores
                     will be included in a future SUF
                release for Starting Cohort 3


==================================================
* Changes introduced to NEPS:SC3 by version 7.0.0 *
==================================================

xTargetCompetencies:
        - items from the competency assessment for domain "maths" in wave 5 had not
            been delivered by the responsible data
                curators in due time for release 5.0.0 through 6.0.1; they have been
                    integrated into the 7.0.0 release


==================================================
* Changes introduced to NEPS:SC3 by version 6.0.1 *
==================================================

Weights:
        - survey design weights (w_t_cal) erroneously had been calibrated to the
            population totals of sex, federal state
                and school type in school year 2010/2011 for version 6.0.0; this only
                    should have happened for the weight
                of participation in wave 1 (w_t1_cal); this eventually led to all cross
                    -sectional and longitudinal weights
                being based on the uncalibrated weight (w_t) in version 6.0.0; this has
                     been fixed

xTargetCompetencies:
        - linkage of WLE estimators for domains "maths" (mag*_sc1u), "reading" (reg*
            _sc1u), and "ict" (icg*_sc1u)
                had been errouneous in all previous releases; this has been fixed


==================================================
* Changes introduced to NEPS:SC3 by version 6.0.0 *
==================================================

CohortProfile:
        - a mistake in the variable tx80230 "panel frame" led to erroneously coded
            values in the variable ID_t and the
                generated variable t723080_g1 "current type of school (constructed)";
                    this has been fixed


==================================================
* Changes introduced to NEPS:SC3 by version 5.0.0 *
==================================================

pParent:
        - the generated variables ISCED, CASMIN and YEARS OF EDUCATION for the surveyed
             parent and its partner
                (p731802_g* and p731952_g*) contained wrong values for wave 4 in
                    Version 4.0.0; this has been fixed
```

```
================================================
* Changes introduced to NEPS:SC3 by version 4.0.0 *
================================================

CohortProfile:
        - in version 3.1.0 and 3.0.0, dummy variables indicating availibility of
            context data from headmasters [tx80524],
                class courses [tx80525], german courses [tx80526] and maths courses [
                    tx80527] erroneously had been set
                to 0 for wave 1 data for all students sampled in special educational
                    needs schools or the migrant
                oversampling sub-sample; this has been fixed

pTarget:
        - in version 3.1.0, coded variables containing the target persons', parents' or
             grandparents' "country of birth"
                [t400000_g*,t400070_g*,t400090_g*,t400220_g*,t400240_g*,t400260_g*,
                    t400280_g*]
                suffered a coding error in the pTarget dataset version 3.1.0, leading
                    to wrong values for all variables in wave 1,
                this also affected the calculated migrational background variables [
                    t400500_g*]. This has been fixed.
                In version 3.1.0, the fix can be manually implemented by using wave 1
                    and 2 data can be used from SUF version 2.0.0,
                for instance by merging variables using the following Stata syntax:
                * -------------------------BEGIN Stata---------------------------------
                // adjust this with the file path to your version 2.0.0 pTarget file
                local oldfile Z:/SUF/On-site/SC3/SC3_O_2-0-0/Stata14/SC3_pTarget_O_2
                    -0-0.dta
                // adjust this with the file path to your version 3.0.0 pTarget file
                local newfile Z:/SUF/On-site/SC3/SC3_O_3-0-0/Stata14/SC3_pTarget_O_3
                    -0-0.dta
                // variable list to be merged
                local updatevarlist t400500_g* t400000_g* t400070_g* t400090_g*
                    t400220_g* t400240_g* t400260_g* t400280_g*
                // open 3.0.0 dataset
                use `"`newfile'"' , clear
                // fill in variables from 2.0.0 dataset
                merge 1:1 ID_t wave using `"`oldfile'"' , keepusing(`updatevarlist')
                    update replace nogenerate assert(master match match_conflict)
                // done
                exit 0
                * ---------------------------END Stata---------------------------------

        - the concept of reflecting migrational background in NEPS SUFs has been
            improved in order to also represent migrants in 3.75th generation;
                thus, the older variables on migrational background [t400500_g1,
                    t400500_g2,t400500_g3] in the pTarget dataset have been renamed
                    using
                the "v1" suffix [t400500_g1v1,t400500_g2v1,t400500_g3v1], and the new
                    ones have been introduced

        - data from the questionnaires on spelling (waves 1 and 3) has been removed
            from xTargetCompetencies and integrated into pTarget
                upon request by the responsible item developers

pParent:
        - the concept of reflecting migrational background in NEPS SUFs has been
            improved in order to also represent migrants in 3.75th generation;
                thus, the older variables on migrational background [p400500_g1,
                    p400500_g2,p400500_g3] in the pParent dataset have been renamed
                    using
```

```
                the "v1" suffix [p400500_g1v1,p400500_g2v1,p400500_g3v1], and the new
                    ones have been introduced

ParentMethods:
        - variable "Willingness: panel participation" [px80400] in dataset
            ParentMethods erroneously did not contain value
                labels for values 0 ["Unavailable"] and 1 ["Available"] in version
                    3.1.0; this has been fixed

xTargetCompetencies:
        - maths items that have been developed for graede 5 and are repeatedly measured
                in grade 7 erroneously had been misnamed since versino 3.0.0; variables
                mag5q301_c_sc3g7, mag5d051_c_sc3g7, mag5d052_c_sc3g7, mag5r251_c_sc3g7,
                     mag5v321_c_sc3g7 and mag5r191_c_sc3g7
                should (in accordance with NEPS' naming convetions for test items) have
                     been named
                mag5q301_sc3g7_c, mag5d051_sc3g7_c, mag5d052_sc3g7_c, mag5r251_sc3g7_c,
                     mag5v321_sc3g7_c and mag5r191_sc3g7_c;
                this has been fixed

        - some items from the domain procedural metacognition that had been surveyed in
             wave 3 were missing in versions 3.0.0 and 3.1.0; this has been fixed

        - data from the questionnaires on spelling (waves 1 and 3) has been removed
            from xTargetCompetencies and integrated into pTarget
                upon request by the responsible item developers

================================================
* Changes introduced to NEPS:SC3 by version 3.1.0 *
================================================

pTarget:
        - variables t31135a and t31035a suffered from a coding error in wave 1 & wave 2
             in version 3.0.0; this has been fixed
                In version 3.0.0, the fix can be manually implemented using the
                    following Stata syntax:
                * -------------------------BEGIN Stata--------------------------------
                // adjust this with the file path to your version 3.0.0 pTarget file
                use "Z:\SUF\On-site\SC3\SC3_O_3-0-0\Stata14\SC3_pTarget_O_3-0-0.dta" ,
                    clear
                local recodevars t31135a t31035a
                foreach var of local recodevars {
                        recode 'var' (4=1)(1=2)(2=3)(3=4) if wave ==1, copyrest
                        recode 'var' (1=4)(2=1)(3=2)(4=3) if wave ==2, copyrest
                }
                save, replace
                exit 0
                * ----------------------------END Stata---------------------------------

        - in version 3.0.0, coded variables containing the target persons', parents' or
             grandparents' "country of birth"
                [t400000_g*,t400070_g*,t400090_g*,t400220_g*,t400240_g*,t400260_g*,
                    t400280_g*]
                suffer a coding error in the pTarget dataset version 3.0.0, leading to
                    "Germany" erroneously being coded as "Afghanistan";
                this also affects the calculated migrational background variables [
                    t400500_g*].
                This has been fixed
                In version 3.0.0, the fix can be manually implemented by using wave 1
                    and 2 data can be used from SUF version 2.0.0,
                for instance by merging variables using the following Stata syntax:
                * -------------------------BEGIN Stata--------------------------------
```

```
                // adjust this with the file path to your version 2.0.0 pTarget file
                local oldfile Z:/SUF/On-site/SC3/SC3_O_2-0-0/Stata14/SC3_pTarget_O_2
                    -0-0.dta
                // adjust this with the file path to your version 3.0.0 pTarget file
                local newfile Z:/SUF/On-site/SC3/SC3_O_3-0-0/Stata14/SC3_pTarget_O_3
                    -0-0.dta
                // variable list to be merged
                local updatevarlist t400500_g* t400000_g* t400070_g* t400090_g*
                    t400220_g* t400240_g* t400260_g* t400280_g*
                // open 3.0.0 dataset
                use `""`newfile'""' , clear
                // fill in variables from 2.0.0 dataset
                merge 1:1 ID_t wave using `""`oldfile'""' , keepusing(`updatevarlist')
                    update replace nogenerate assert(master match match_conflict)
                // done
                exit 0
                * ------------------------------END Stata---------------------------------
        - Variable [t27111d_O] has been completely removed

CohortProfile:
        - an indicator containing the field of determined special educational needs
            from the list of students has been added
                [tx80505_D] is a dummy indicating special educational needs
                [tx80505_R] describes the need more specific (only available in
                    RemoteNEPS or via on-site access)

        - two indicators containing the region [tx80109_g1] and the German federal
            state [tx80109_g2R] of the sampled institution have been added;
                the latter is only available in RemoteNEPS or via on-site access

xTargetCompetencies:
        - in the xTargetCompetencies data set file, items from the orthography testing
            in wave 3 were missing;
                this has been fixed, including updated values of wave 1 orthography
                    test items

spSiblings:
        - in the spSiblings data set, system missing values in variables "'Siblings
            date of birth - month" [p73221m] and
                "'Siblings date of birth - year" [p73221y] are incorrectly coded in
                    version 3.0.0, leading to implausible birth dates;
                this has been fixed.

pParent:
        - the variables [p727001] and [p727002] (formerly anonymized because they
            contain federal states)
                are now available with full information in RemoteNEPS or via on-site
                    access
        - variable [t751016] now has been correctly renamed to [p751016]
        - variable [askl] now has been correctly renamed to [p723400]


================================================
* Changes introduced to NEPS:SC3 by version 3.0.0 *
================================================

General:
        - starting with this release, all NEPS Scientific Use Files will ship with an
            additional, unicode-enabled Stata data set version;
                this version is only readable in Stata version 14 or younger, and is
                    placed in the subdirectory "Stata14"
```

```
        - translation for all meta data (variable and value labels, question texts, etc
          ) have been revised and completed
        - meta data for all variables have been revised and updated where appropriate
        - additional wave 3 has been incorporated into the data, including observations
            from a sample refreshment in wave 3
        - regional information for German federal states have been added (for waves 1
          through 3, restrospectively) to the download SUF,
              more fine-grained information to the RemoteNEPS and onsite variant
        - information from a mid-year status update mailing regarding the school status
            of target persons in the individual field have been
              incorporated into CohortProfile


pParent:
        - forwarded general and vocational educational information for parent and
          partner in time to construct more reliable ISCED-97 and CASMIN scores
        - ISCED-97: added code "4A" for parent and partner when reporting both "Abitur"
            etc.  and "Vocational Training" (not university in this context)
        - minor changes to achieve more precise ISCED-97 and CASMIN values für parents
          and partner

pTarget:
        - variables "idealistic educational aspiration" [t31035a] and "realistic
          educational aspiration" [t31135a] suffered an encoding
              error for wave 1 data in version 2.0.0; this has been fixed; a
                  temporary workaround can be achieved by using version 2.0.0
              data sets and correctly recode those values using the following Stata
                  syntax snippet:
              . recode t31035a t31135a (4=1) (1=2) (2=3) (3=4) if wave==1 //
        - variables for persons who did not respond to a part of the questionnaire in
          special needs schools erroneously
              contained the value "missing by design" (-54) instead of "not
                  participated" (-56) in the pTarget data set in version 2.0.0;
              this has been fixed
        - when generating variable "Global self-esteem" [t66003a_g1] in the pTarget
          data set, variable "Global self-esteem: competence" [t66003d]
              erroneously had been ignored in version 2.0.0;
              t66003a_g1 could temporarily be re-generated from version 2.0.0 using
                  the following Stata syntax:
              * -------------------------BEGIN Stata--------------------------------
              nepsmiss t66003a t66003b t66003c t66003d t66003e t66003f t66003g
                  t66003h t66003i t66003j
              tempvar t66003b_r t66003e_r t66003f_r t66003h_r t66003i_r rowmissings
              recode t66003b (1=5) (2=4) (3=3) (4=2) (5=1), generate('t66003b_r')
              recode t66003e (1=5) (2=4) (3=3) (4=2) (5=1), generate('t66003e_r')
              recode t66003f (1=5) (2=4) (3=3) (4=2) (5=1), generate('t66003f_r')
              recode t66003h (1=5) (2=4) (3=3) (4=2) (5=1), generate('t66003h_r')
              recode t66003i (1=5) (2=4) (3=3) (4=2) (5=1), generate('t66003i_r')
              egen 'rowmissings'=rowmiss(t66003a 't66003b_r' t66003c t66003d ///
              't66003e_r' 't66003f_r' t66003g 't66003h_r' 't66003i_r' t66003j)
              egen 'target_variable'=rowtotal(t66003a 't66003b_r' t66003c t66003d ///
              't66003e_r' 't66003f_r' t66003g 't66003h_r' 't66003i_r' t66003j) if '
                  rowmissings'==0 & wave==3
              replace 'target_variable'=-54 if wave!=3
              label variable 'target_variable' "Global self-esteem"
              replace 'target_variable'=-55 if missing('target_variable')
              * --------------------------END Stata---------------------------------
              this issue has been fixed;

spParentSchool:
        - for the sake of consistency between NEPS Starting Cohorts, the data set "
          spSchool" has been renamed to "spParentSchool"
```

```
spParentGap:
        - for the sake of consistency between NEPS Starting Cohorts, the data set "
          spGap" has been renamed to "spParentGap"

================================================
* Changes introduced to NEPS:SC3 by version 2.0.0 *
================================================

General:
        - SPSS data sets now ship with the same "VARIABLE ATTRIBUTES" as Stata data
          sets' "characteristics"
        - metadata for all datasets has been revised and updated where appropriate
        - variables now ship with a characteristic 'NEPS_instname' attached in Stata
          datasets, reporting the variable name used in the survey
        - wave 2 data has been fully integrated into the data
        - a new dataset "Weights" has been added, reflecting panel weights for the
          cohort; documentation is available online
        - a new dataset "spSibling" has been added, reflecting the target person's
          siblings reported in the parent's interview
        - a new dataset "pTargetMicrom" has been added for onsite access, reflecting
          spatial data from "microm Micromarketing-Systeme und Consult GmbH"
        - a new dataset "pInstitutionMicrom" has been added for onsite access,
          reflecting spatial data from "microm Micromarketing-Systeme und Consult
          GmbH"
        - several bugfixes and enhancements have been integrated into this new release,
           influencing various variables;
              only the most important ones are listed in this change log

pParent:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xParent" to "pParent"
        - the interview process in parent's interviews does not guarantee unique ids
          for parents;
              thus, the identifier in this dataset is no longer "ID_p", but the
                  target person's "ID_t"
        - three variables with information about the target person's migrational status
           have been calculated [p400500_g1, p400500_g2, p400500_g3];
              a working paper on the generation process and theoretical background is
                  forthcoming
        - values 5, 6 and 7 have been recoded to 7, 8 and 9 in 'Highest education
          qualification (ISCED)' [p731802_g1]
        - values 5, 6 and 7 have been recoded to 7, 8 and 9 in 'Partner: Highest
          education qualification (ISCED)' [p731852_g1]
        - value 96 has been recoded to -20 in 'Partner: (Highest) vocational education
          certificate' [p731863] in accordance with official NEPS missing codes
        - EGP generation syntax was adjusted due to errors in the derivation syntax (
          particularly classes IVc and V) [p731904_g8, p731954_g8]
        - German EGP value labels have been corrected [p731904_g8, p731954_g8]
        - CASMIN [p731802_g2 & p731852_g2]: Class assignment slightly modified
        - ISCED [p731802_g1 & p731852_g1]: Civil servants of the medium grade are now
          identifiable
        - 'SDQ-Scale: Prosocial behaviour' [p67801a_g1] has been corrected to only
          contain a sum score if all included items are non-missing

CohortProfile:
        - older weighting variables ('Standardized design weight' [weight_design_std]
          and 'Design weight' [weight_design]) are now
              deprecated and have been removed
        - the interview process in parent's interviews does not guarantee unique ids
          for parents;
```

```
                as "ID_p" therefore has been replaced by "ID_t" and is no longer needed
                    to link datasets, it has been removed from CohortProfile
        - variable 'Test: survey day (month)' [test] has been renamed to [testm]
        - variable 'Test: survey day (year)' [test] has been renamed to [testy]

pEducator:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xEducator" to "pEducator"

pInstitution:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xInstitution" to "pInstitution"

pCourseClass:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xCourseClass" to "pCourseClass"

pCourseGerman:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xCourseGerman" to "pCourseGerman"

pCourseMath:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xCourseMath" to "pCourseMath"

pTarget:
        - as wave 2 data makes this a panel dataset, the filename has changed from "
          xTarget" to "pTarget"
        - three variables with information about the target person's migrational status
            have been calculated [t400500_g1, t400500_g2, t400500_g3];
              a working paper on the generation process and theoretical background is
                  forthcoming
        - 'SDQ-Scale: prosocial behaviour' [t67801a_g1] has been corrected to only
          contain a sum score if all included items are non-missing
        - the scale of variable 'Helps other voluntarily' [t67801i] has been
          erroneously reversed in generation of
              'SDQ-Scale: prosocial behaviour' [t67801a_g1]; this has been fixed

spGap:
        - the interview process in parent's interviews does not guarantee unique ids
          for parents;
              thus, the identifier in this dataset is no longer "ID_p", but the
                  target person's "ID_t"

spSchool:
        - the interview process in parent's interviews does not guarantee unique ids
          for parents;
              thus, the identifier in this dataset is no longer "ID_p", but the
                  target person's "ID_t"
```