NEPS
National Educational Panel Study

Research Data

FDZ-LIfBi

Data Manual

NEPS Starting Cohort 5—First-Year Students

From Higher Education to the Labor Market

Scientific Use File Version 18.0.0

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Copyrighted Material
Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3, 96047 Bamberg
Director: Prof. Dr. Cordula Artelt
Administrative Director: Dr. Stefan Echinger
Bamberg; June 14, 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 5–First-Year Students, From Higher Education to the Labor Market, Scientific Use File Version 18.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

This data manual for Starting Cohort 5–First-Year Students "From Higher Education to the Labor Market" 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

For their support in writing this manual, special thanks go to DZHW Hannover:
Isabelle Fiedler, Stefanie Gäckle, Annika Grieb, Marie Kühn, Uta Liebeskind, Katrin Mergard, Andreas Ortenburger, Hilde Schaeper

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**
1.1 About this manual . . . 1
1.2 Further documentation . . . 1
1.3 Data release strategy . . . 3
1.4 Data access . . . 5
1.5 Publications with NEPS data . . . 6
1.6 Rules and recommendations . . . 7
1.7 On using the Federal State label *(Bundeslandkennung)* . . . 9
1.8 User services . . . 9
1.9 Contacting the Research Data Center . . . 11

2 Sampling and Survey Overview **12**
2.1 From higher education to the labor market . . . 12
2.2 Sampling strategy . . . 13
2.3 Competence measures . . . 14
2.4 Survey overview and sample development . . . 17
2.4.1 Wave 1: 2010/2011 (CATI+competencies) . . . 18
2.4.2 Wave 2: 2011 (CAWI) . . . 19
2.4.3 Wave 3: 2012 (CATI) . . . 20
2.4.4 Wave 4: 2012 (CAWI) . . . 21
2.4.5 Wave 5: 2013 (CATI+competencies) . . . 22
2.4.6 Wave 6: 2013 (CAWI) . . . 23
2.4.7 Wave 7: 2014 (CATI+competences) . . . 24
2.4.8 Wave 8: 2014 (CAWI) . . . 25
2.4.9 Wave 9: 2015 (CATI) . . . 26
2.4.10 Wave 10: 2016 (CATI) . . . 27
2.4.11 Wave 11: 2016 (CAWI) . . . 28
2.4.12 Wave 12: 2017 (CATI) . . . 29
2.4.13 Wave 13: 2018 (CATI) . . . 30
2.4.14 Wave 14: 2018 (CAWI) . . . 31
2.4.15 Wave 15: 2019 (CATI) . . . 32
2.4.16 Wave 16: 2020 (CATI) . . . 33
2.4.17 Wave 17: 2020 (CAWI) . . . 34
2.4.18 Wave 18: 2021 (CATI) . . . 35

3 General Conventions **36**
3.1 File names . . . 36
3.2 Variables . . . 38
3.2.1 Conventions for general variable naming . . . 38
3.2.2 Conventions for competence variable naming . . . 41
3.2.3 Labels . . . 44

3.3 Missing values . . . 45
3.4 Generated variables . . . 48

4 Data Structure **50**
4.1 Overview . . . 50
4.2 Identifiers . . . 51
4.3 Panel data . . . 51
4.4 Episode or spell data . . . 52
4.4.1 Edition of the life course . . . 53
4.4.2 Revoked episodes . . . 55
4.4.3 Subspells and harmonization of episodes . . . 55
4.5 Data files . . . 61
4.5.1 Basics . . . 63
4.5.2 Biography . . . 65
4.5.3 CohortProfile . . . 67
4.5.4 EditionBackups . . . 69
4.5.5 Education . . . 71
4.5.6 MethodsCATI . . . 73
4.5.7 MethodsCAWI . . . 75
4.5.8 MethodsCompetencies . . . 77
4.5.9 pTargetCATI . . . 79
4.5.10 pTargetCAWI . . . 81
4.5.11 pTargetMicrom . . . 83
4.5.12 spChild . . . 85
4.5.13 spChildCohab . . . 87
4.5.14 spCourses . . . 89
4.5.15 spEmp . . . 91
4.5.16 spFurtherEdu1 . . . 93
4.5.17 spFurtherEdu2 . . . 95
4.5.18 spGap . . . 97
4.5.19 spInternship . . . 99
4.5.20 spMilitary . . . 101
4.5.21 spParLeave . . . 103
4.5.22 spPartner . . . 105
4.5.23 spSchool . . . 107
4.5.24 spSchoolExtExam . . . 109
4.5.25 spSibling . . . 111
4.5.26 spUnemp . . . 113
4.5.27 spVocBreaks . . . 115
4.5.28 spVocExtExam . . . 117
4.5.29 spVocPrep . . . 119
4.5.30 spVocTrain . . . 121
4.5.31 StudyStates . . . 124
4.5.32 Weights . . . 126

4.5.33 xEcoCAPI . . . 128
4.5.34 xInstitution . . . 130
4.5.35 xPlausibleValues . . . 132
4.5.36 xTargetCompetencies . . . 134
4.5.37 xTargetCORONA . . . 136

5 Special Issues **138**
5.1 Special Types of Variables . . . 138
5.1.1 Service Variables . . . 138
5.1.2 Auxiliary variables . . . 139
5.1.3 Version variables . . . 139
5.1.4 Preload Variables . . . 140
5.2 Coding field of study . . . 140
5.2.1 Recruitment . . . 140
5.2.2 Panel Waves . . . 141
5.3 Coding of Higher Education Institutions . . . 142
5.4 Special features of interruption episodes in spVocTrain . . . 143
5.5 Teacher Education Students and Teachers . . . 143
5.6 Wave Specific Issues . . . 146

A References **147**

B Appendix **149**
B.1 R examples . . . 149
B.2 Release notes . . . 179
B.3 Comparison of _v1 variables . . . 189

1 Introduction

1.1 About this manual

This manual is intended to facilitate your work with data of NEPS Starting Cohort 5–First-Year Students (NEPS SC5). 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 5 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 5. 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 First-Year Students > Documentation

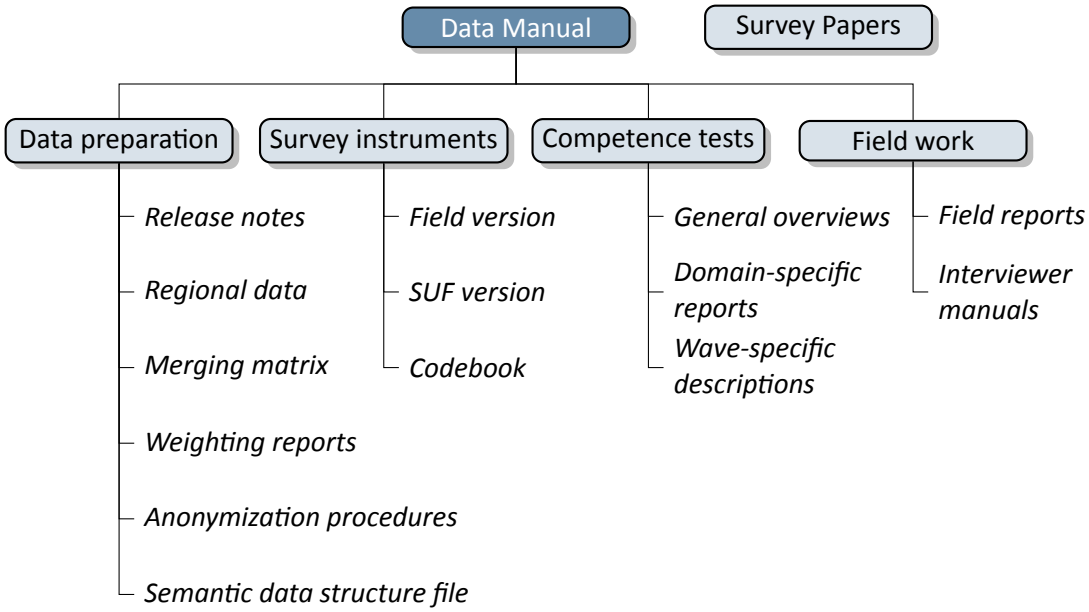

Figure 1: NEPS supplementary data documentation

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 (not for Starting Cohort 1).

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:SC5:18.0.0`. Other releases of Scientific Use Files for Starting Cohort 5 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 5

SUF Version	DOI	Date of release
18.0.0 (current)	doi:10.5157/NEPS:SC5:18.0.0	**2023-06-13**
17.0.0	doi:10.5157/NEPS:SC5:17.0.0	2022-11-07
16.0.0	doi:10.5157/NEPS:SC5:16.0.0	2022-05-16
15.0.0	doi:10.5157/NEPS:SC5:15.0.0	2021-05-19
14.1.0	doi:10.5157/NEPS:SC5:14.1.0	2020-12-02
14.0.0	doi:10.5157/NEPS:SC5:14.0.0	2020-05-27
13.0.0	doi:10.5157/NEPS:SC5:13.0.0	2020-02-14
12.0.0	doi:10.5157/NEPS:SC5:12.0.0	2019-07-26
11.0.0	doi:10.5157/NEPS:SC5:11.0.0	2018-09-06
10.0.0	doi:10.5157/NEPS:SC5:10.0.0	2018-04-19
9.0.0	doi:10.5157/NEPS:SC5:9.0.0	2017-06-23
8.0.0	doi:10.5157/NEPS:SC5:8.0.0	2016-12-23
6.0.0	doi:10.5157/NEPS:SC5:6.0.0	2016-03-31
4.0.0	doi:10.5157/NEPS:SC5:4.0.0	2014-09-30

(...)

Table 1: (continued)

SUF Version	DOI	Date of release
3.1.0	doi:10.5157/NEPS:SC5:3.1.0	2014-05-16
3.0.0	doi:10.5157/NEPS:SC5:3.0.0	2013-07-05

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 (`pInstitution`) 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 5.

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. (2023). *National Educational Panel Study, Scientific Use File of Starting Cohort First-Year Students*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC5:18.0.0

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 5–First-Year Students, From Higher Education to the Labor Market, Scientific Use File Version 18.0.0*. 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 5: First-Year Students (SC5), Wave 18, Questionnaires (SUF Version 18.0.0)*. 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 5.

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 From higher education to the labor market

German higher education system has been facing a number of challenges and developments since the early 2000ies, that raised new issues for research. To name but a few, there is the introduction of a two-stage structure in higher education according to the Bologna Process, a growing demand for outcome orientation, the evolution of higher education towards lifelong learning, an increase of (international) competitiveness, and the emerging shortage of highly qualified professionals. At the same time, key issues remained core challenges for the higher education system, such as student dropouts, social selectivity in university entrance, and the relationship between higher education and working life. In order to answer research questions associated with these issues, a cohort of first-year students was followed through their years of study since winter term 2010/11, including their entrance into working life. Central issues to be studied are educational choices, the outcomes of university education, and the entry into the job market.

The main focus is on

- Educational choices during the course of studies and success in studies: What are the determinants of educational decisions and success in studies while studying at a higher education institution – such as dropping out, changing subjects, studying abroad, and pursuing a Master's degree? What is the importance of competencies and social factors, such as social background, gender or migration experiences in this process? Which consequences do decisions have for subsequent education and working life?
- Entrance into working life and professional success: When thinking about students' transition into the job market and their professional success (e.g., occupational position, income, employment security), how important are acquired competencies, on the one hand based on formal qualifications (diplomas), social background, gender, and on the other hand based on social and cultural capital? What role do general competencies play in comparison to subject-specific ones?
- Students' competencies: Which general competencies do students possess to crucial points of time in their students' and young adults' lifecourse (beginning of studies, end of studies/ labour market entry)? How does the competence level influence transitions during studies and beyond (change of subject, higher education drop out, transition to the labour market)? How do competencies correlate with learning environments provided by higher education institutions?

2.2 Sampling strategy

The target population of Starting Cohort 5 is defined as all first-year students of the academic year 2010/2011, independent of their nationality and their knowledge of the German language, who are:

- enrolled for the first time in a public or state-approved institution of higher education in Germany
- aiming at a Bachelor's degree or a state examination (Staatsexamen) in medicine, law, pharmacy, and teaching, or a diploma or Master's degree in Roman Catholic or Protestant theology or specific art and design degrees
- not attending higher education institutions run by Federal Ministries or Federal States for members of their public services (e. g., University of Applied Labour Studies/University of the German Federal Armed Forces Munich/Universität der Bundeswehr München)

The sampling process was designed to incorporate an oversampling of teacher education students and students at private higher education institutions. For that reason, a stratified cluster approach has been applied. Administrative data provided by the Federal Statistical Office of Germany constituted the corresponding sampling frame. Each cluster referred to the total of students enrolled in a certain subject at a particular higher education institution (e. g., social sciences at the University of Bamberg). On the primary level, the stratification differentiated between the following four strata; on the secondary level these strata were combined with groups of related subjects:

- clusters linked to teacher education at public universities
- clusters linked to all other fields of study at public universities
- clusters linked to all fields of studies at public universities of applied sciences (Fachhochschulen)
- clusters linked to all degree programs at private higher education institutions

In a second step, all institutions of selected clusters were contacted by the German Centre of Higher Education Research and Science Studies (DZHW) in order to gain access to the students. The administration of 261 institutions declared their cooperativeness, thereof 104 public universities, 108 public universities of applied sciences, and 49 private university institutions.

In the subsequent recruitment process, two different modes of contact were employed to approach the students and to receive their consent to participate in the panel study:

- conventional mail via higher education institutions administration
- personal information in lectures for freshmen students in the selected fields of studies via interviewers

The former strategy has been applied at all sampled institutions. Recruiting questionnaires in prepared envelopes were transferred to the university administrations together with detailed instructions on how to select the targeted student population. Part of this instruction was the request to include all non-traditional first-year students, i. e., all students with a higher education admission other than the general higher education certificate (Abitur or Fachabitur). It was the task of the higher education institution to compile the respective postal addresses and to send the letters plus reminder letters. Altogether 16,887 filled questionnaires were sent back to the survey agency. The latter strategy presupposed the explicit agreement by the higher education institution and the lecturer to recruit students in appropriate freshmen courses by professional interviewers. In the course of 299 visits at 99 higher education institutions, another 17,229 filled questionnaires could be collected. While the two strategies were conducted parallel during the winter semester 2010/2011, a simplified procedure was applied in the summer semester 2011. Based on postal distribution and display of reduced questionnaires, so-called NEPS address cards, additional 4,169 contact information were gathered.

The returned information of all 38,285 persons were then checked with regard to the belonging to the target population, the existence of double recruitments, and the quality of provided contact details. Finally, 21,438 cases were administrated in the first CATI survey wave of Starting Cohort 5. This first CATI was the prerequisite for staying in the panel.

The sampling design and its consequences for the derivation of sampling weights are fully described in Zinn et al., 2017. Further remarks on the recruiting process are given in the CATI field report of the first survey wave (in German only). Both documents are available on our website at:

→ `www.neps-data.de > Data Center > Data and Documenation > Starting Cohort First-Year Students > Documentation`

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.[1] As a consequence, variable names for

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

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).

The next table shows the schedule of competence measures in Starting Cohort 5 with domains by waves and test modus.

Table 2: Schedule of competence measures. P = Paper-Based Test (proctored), C = Computer-Based Test (proctored), W = Web-Based Test (unproctored)

		2011 **Wave 1** (2nd Sem.)	2013 **Wave 5** (6th Sem.)	2014 **Wave 7** (7th Sem.)	2017 **Wave 12** (13th Sem.)[3]
Domain-General Competencies					
DGCF: Cognitive Basic Skills	dg	—	P, C, W	—	—
Domain-Specific Competencies					
Reading Competence[1]	re	P	—	—	C, W
Reading Speed	rs	P	—	—	—
Mathematical Competence[1]	ma	P	—	—	C, W
Scientific Competence[1]	sc	—	P, C, W	—	—
Metacompetencies					
ICT Literacy[1]	ic	—	P, C, W	—	—
Stage-Specific Competencies					
Business Administration and Economics[2]	ba	—	—	P	—
English Reading Competence[1]	ef	—	—	—	C, W

[1] Subsequent to the respective competence test the target persons had to assess their own test performance (Procedural Metacognition, mp).
[2] Reduced testing: In wave 7, the stage-specific competence test (ba) was realized in a subsample of students and graduates of business sciences only.
[3] Reduced testing: In wave 12, a randomized allocation of competence tests with two out of the three domains (re, ma or re, ef or ma, ef) has been applied.

2.4 Survey overview and sample development

This section informs about the progress of the Starting Cohort 5 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 First-Year Students > Documentation

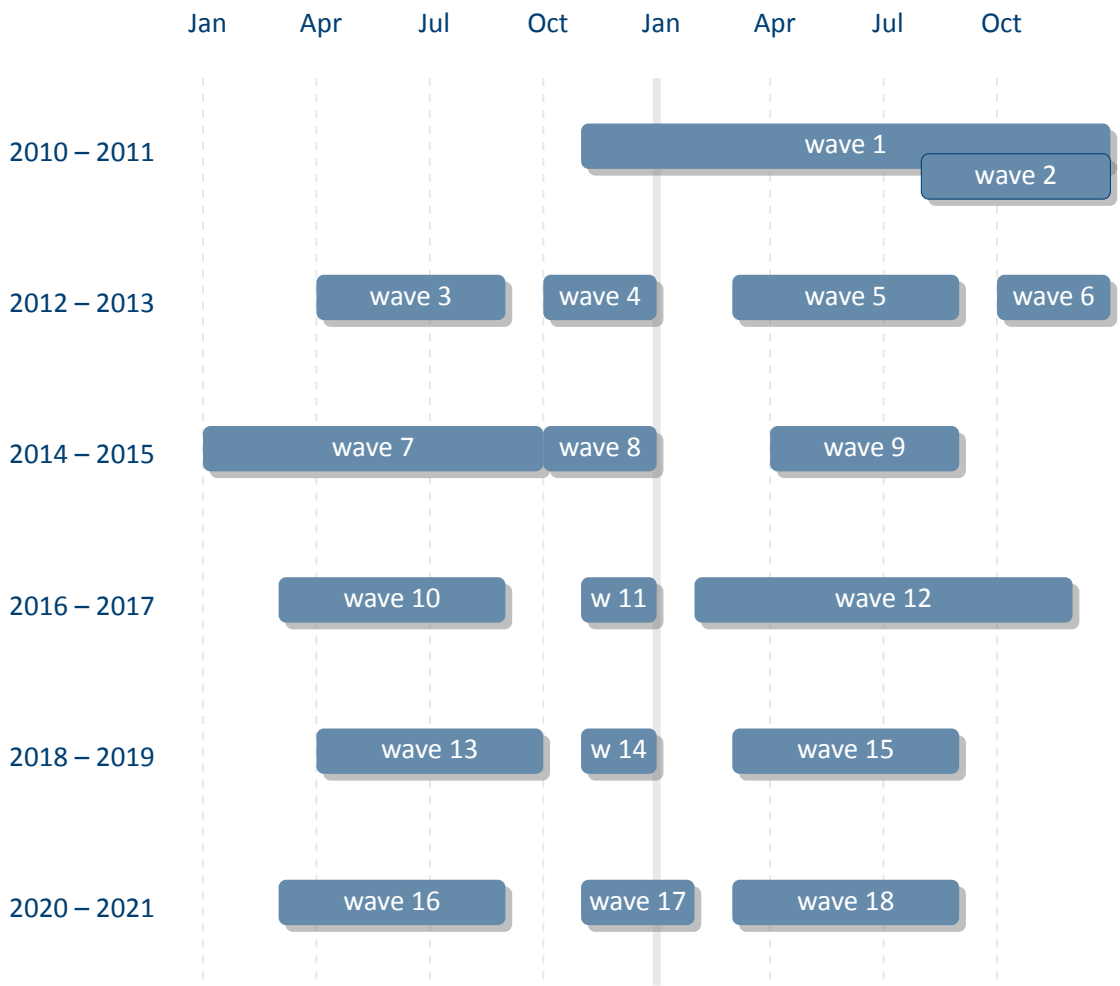

Figure 2: Panel progress of Starting Cohort 5

2.4.1 Wave 1: 2010/2011 (CATI+competencies)

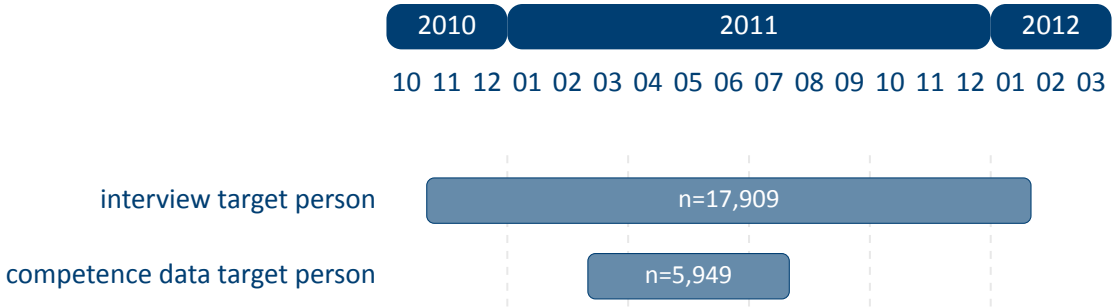

Figure 3: Field times and realized case numbers in wave 1

- Target persons

  Sample First-year students in winter semester 2010/11 (for details about the sampling strategy, see section 2.2)

  Competence tests Reading Competence, Reading Speed, Mathematical Competencies

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Written questionnaires (in each case for recruiting and competence test, PAPI) and computer-assisted telephone interview (CATI)

2.4.2 Wave 2: 2011 (CAWI)

Figure 4: Field times and realized case numbers in wave 2

- Target persons

  Sample Participants of the first wave willing to take part in the panel

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.3 Wave 3: 2012 (CATI)

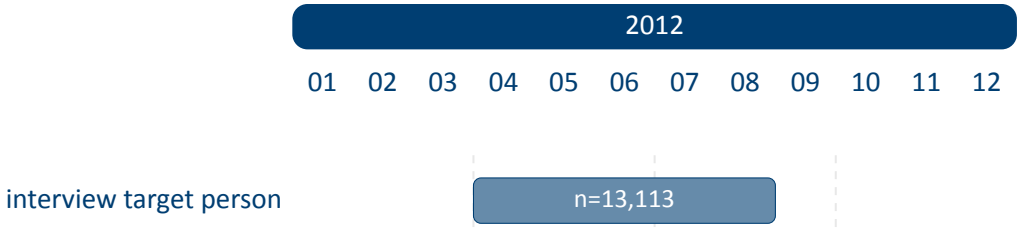

Figure 5: Field times and realized case numbers in wave 3

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.4 Wave 4: 2012 (CAWI)

Figure 6: Field times and realized case numbers in wave 4

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.5 Wave 5: 2013 (CATI+competencies)

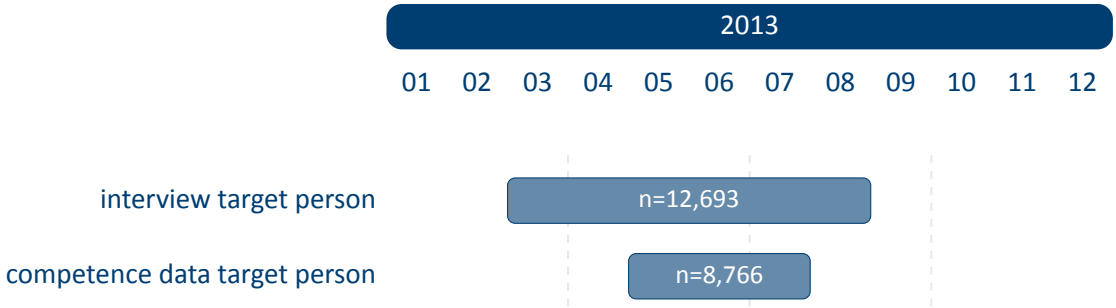

Figure 7: Field times and realized case numbers in wave 5

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Competence tests DGCF (Cognitive Basic Skills), Scientific Competence, ICT Literacy

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI) and group testing (conventional paper-based testing (PAPI), paper-based testing with electronic pens (E-Pen) or computer-based testing with notebooks (CBA)) or individual testing (computer-based online testing, CBWA)

2.4.6 Wave 6: 2013 (CAWI)

Figure 8: Field times and realized case numbers in wave 6

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.7 Wave 7: 2014 (CATI+competences)

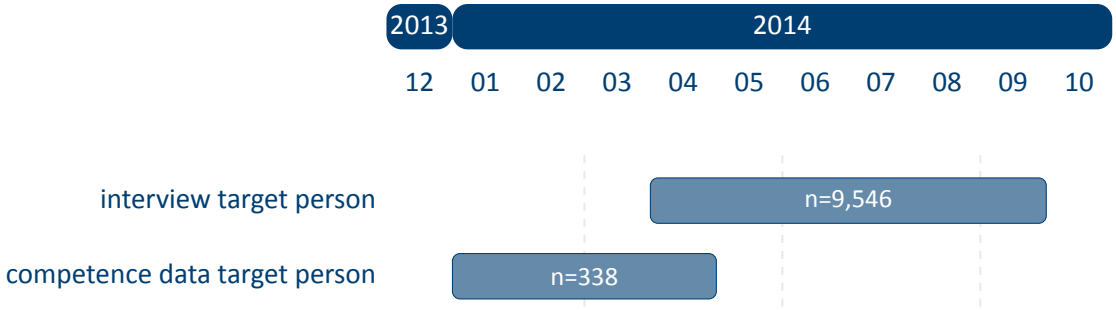

Figure 9: Field times and realized case numbers in wave 7

- Target persons (Subsample A)

  Current wave All students excluding the teaching-oversampling. (see section 2.2 for more information about this subpopulation).

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

- Target persons (Subsample B)

  Current wave Selected students who study an economic subject or have graduated from such studies. (identifiable via `tx80921` in `CohortProfile`).

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Competence tests Business Administration and Economics

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Paper-based competence testing within a personal-verbal interview (CAPI)

2.4.8 Wave 8: 2014 (CAWI)

Figure 10: Field times and realized case numbers in wave 8

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.9 Wave 9: 2015 (CATI)

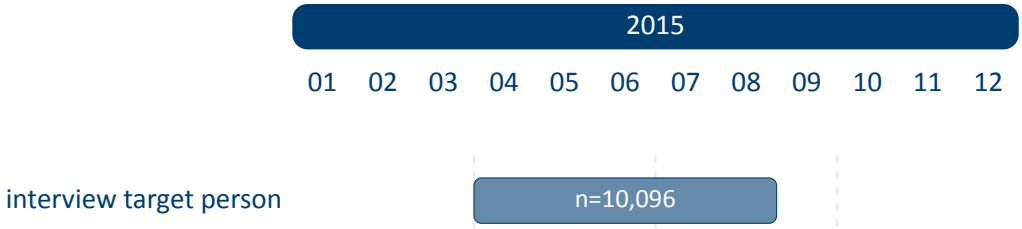

Figure 11: Field times and realized case numbers in wave 9

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.10 Wave 10: 2016 (CATI)

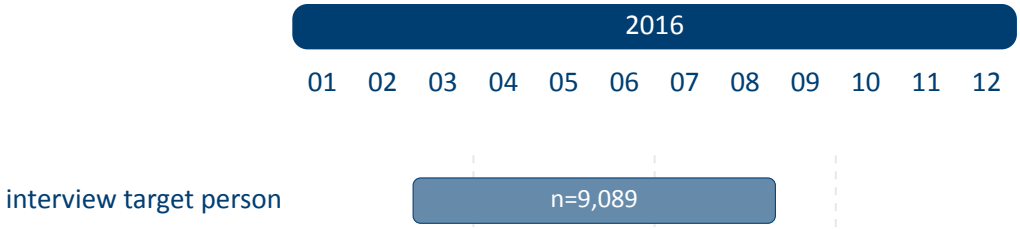

Figure 12: Field times and realized case numbers in wave 10

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.11 Wave 11: 2016 (CAWI)

Figure 13: Field times and realized case numbers in wave 11

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas – Institute for Applied Social Sciences, Bonn

  Mode of survey Online survey (CAWI)

2.4.12 Wave 12: 2017 (CATI)

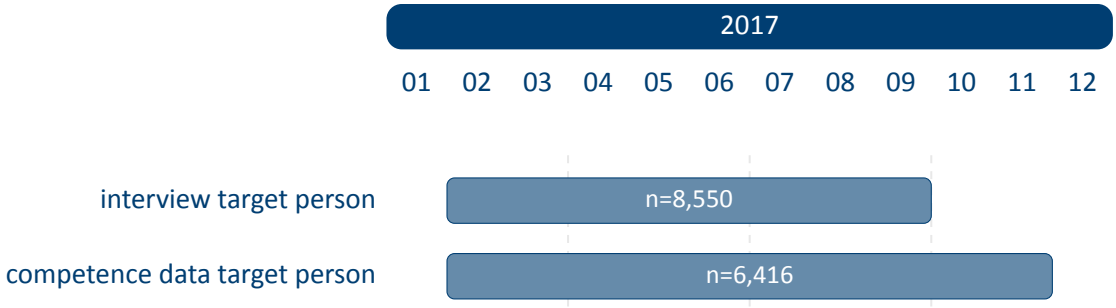

Figure 14: Field times and realized case numbers in wave 12

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Competence tests Reading Competence, Mathematical Competence, English Reading Competence

  Data collection infas - Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI) and group testing (computer-based testing with notebooks (CBA)) or individual testing (computer-based online testing, CBWA)

2.4.13 Wave 13: 2018 (CATI)

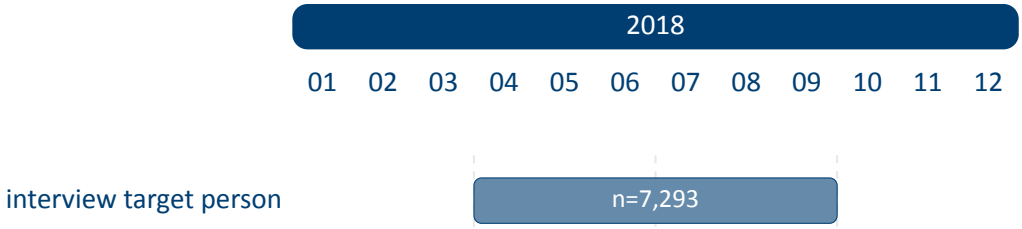

Figure 15: Field times and realized case numbers in wave 13

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas - Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.14 Wave 14: 2018 (CAWI)

Figure 16: Field times and realized case numbers in wave 14

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.15 Wave 15: 2019 (CATI)

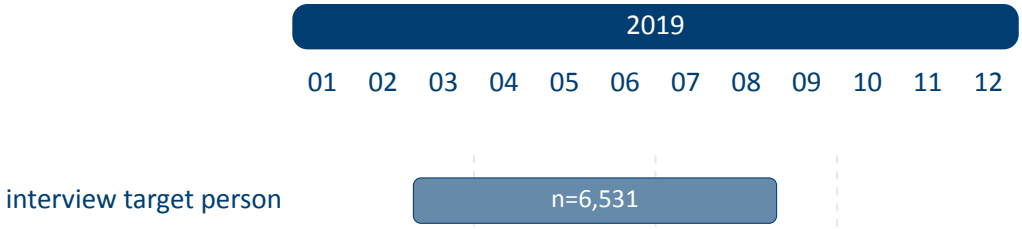

Figure 17: Field times and realized case numbers in wave 15

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas - Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.16 Wave 16: 2020 (CATI)

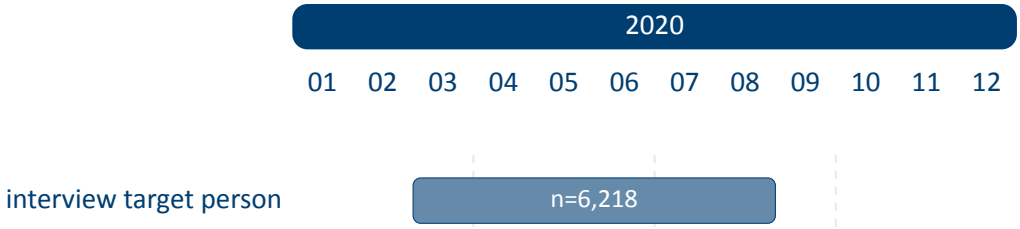

Figure 18: Field times and realized case numbers in wave 16

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas - Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

2.4.17 Wave 17: 2020 (CAWI)

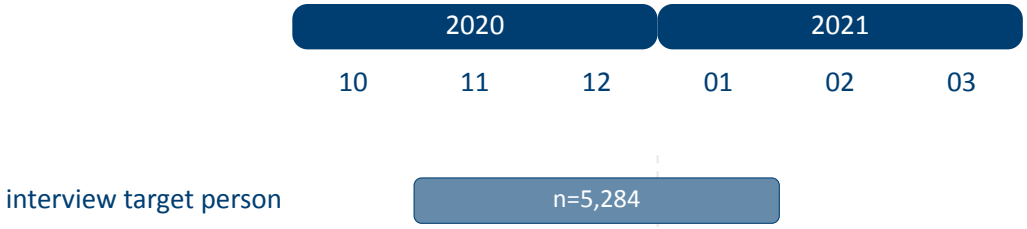

Figure 19: Field times and realized case numbers in wave 17

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection DZHW - German Centre for Higher Education Research and Science Studies, Hannover

  Mode of survey Online survey (CAWI)

2.4.18 Wave 18: 2021 (CATI)

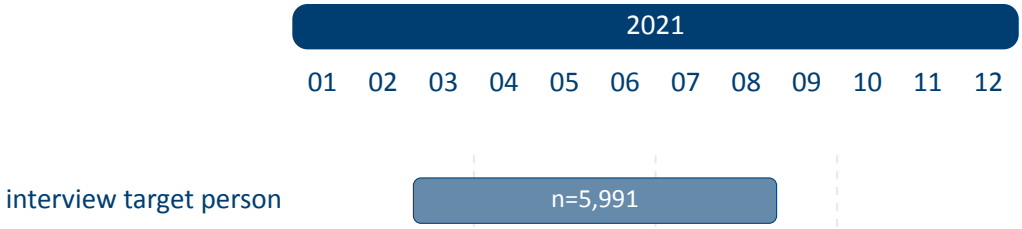

Figure 20: Field times and realized case numbers in wave 18

- Target persons

  Sample Panel sample. Follow-up survey with interviewees willing to participate in the panel.

  Data collection infas - Institute for Applied Social Sciences, Bonn

  Mode of survey Computer-assisted telephone interview (CATI)

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 4.5.4 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 5, 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 `SC5_CohortProfile_D_18.0.0.dta` refers to the *CohortProfile* data of *Starting Cohort 5* in its *Download* version of the Scientific Use File release *18.0.0*.

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 21. 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.

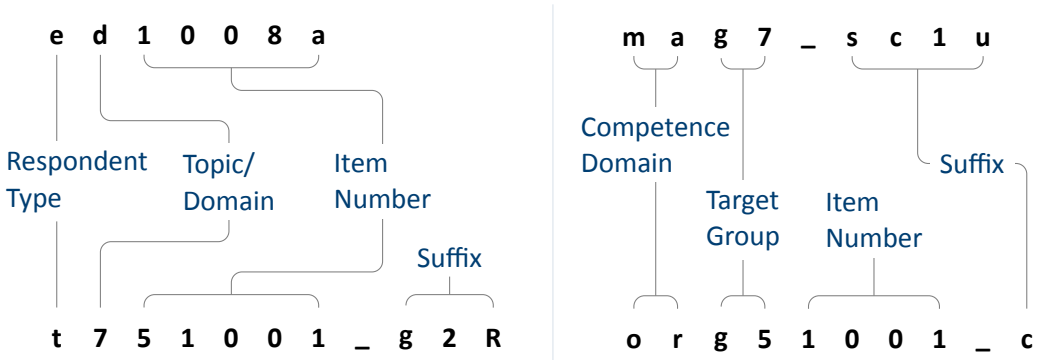

Figure 21: General variable naming (left) and competence variable naming (right)

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. Versionized variables are listed in section B.3.*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`).

Specific variables for (prospective) teachers

Certain parts of the survey in Starting Cohort 5 refer to teaching. The corresponding information in the datasets can be identified by variable names: Variables with the first three characters `tg6` or `tg8` indicate questions specifically addressed to (prospective) teachers.

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.[2]

2 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.

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", 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

(...)

Table 5: (continued)

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

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

(...)

Table 6: (continued)

Code	Meaning	Note
–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)
–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 22 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.

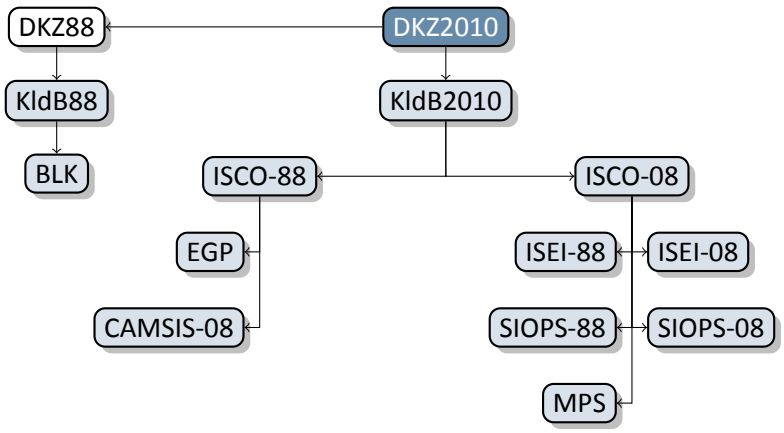

Figure 22: Derivation paths for several occupational scales and schemes provided in the NEPS

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.

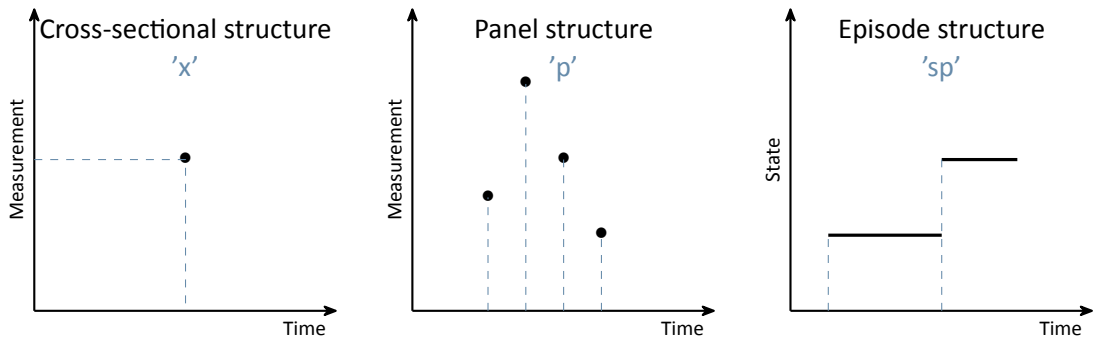

Figure 23: Different types of data structures

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 25).

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 5 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`.

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. 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 .

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.

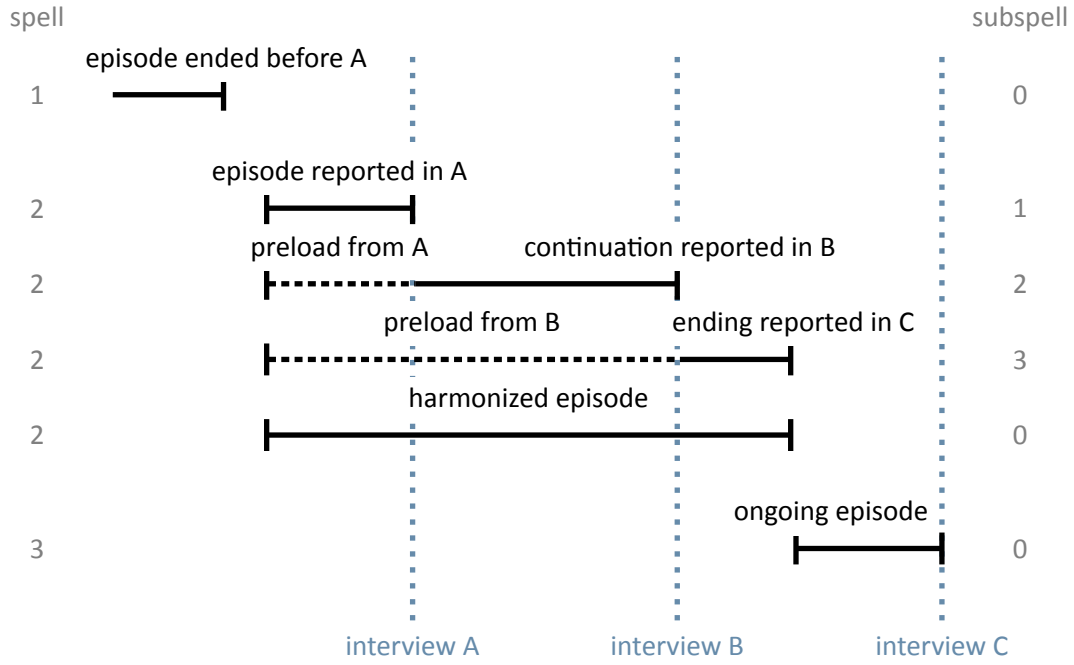

Figure 24: Logic of subspells

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 24), 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.

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 24). 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 24).

The sample episodes from figure 24 correspond to the data structure presented in table 7 *before* any episode harmonization.[3] 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 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,

3 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).

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. 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 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. There are six rules that are applied for selecting the relevant piece of information for the harmonized spell. Which of these rules is used for a variable depends on content-related criteria. Data users can identify the respective rule in the additional attributes or characteristics of each variable:

first_noedit 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). Missing values from -59 to -50 in the first subspell as well as the missing value -29 are **not** transferred to the harmonized spell.[4] In case that there are such missings in the first subspell, the next non-missing value from the subsequent subspells is taken instead.

4 There is one exception: If the missing code -53 (anonymized) is given in the first subspell, this value is copied to the harmonized spell.

last_noedit 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). Missing values from -59 to -50 as well as the missing value -29 in the last subspell are **not** transferred to the harmonized spell.[5] In case that there are such missings in the last subspell, the next non-missing value from the previous subspells is taken instead.

first_noeditnosys The harmonization of most variables follows either the *first_noedit* or the *last_noedit* 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_noedit* 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. Missing values from -59 to -50 as well as the missing value -29 and system missings (.) in the first subspell are **not** transferred to the harmonized spell.

last_noeditnosys 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.[6] 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_noeditnosys* (last valid value found in the subspells of an episode) is used for harmonization. Missing values from -59 to -50 as well as the missing value -29 and system missings (.) in the last subspell are **not** transferred to the harmonized spell.

first_all This rule is identical to *first_noedit* with the exception that **all** missing codes from the first subspell are transferred to the harmonized spell.

last_all This rule is identical to *last_noedit* with the exception that **all** missing codes from the last subspell are transferred to the harmonized spell.

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. The information is stored in the datasets for each relevant variable in the additional attributes or characteristics. The harmonization can also be viewed upon specific request.

5 There is one exception: If the missing code -53 (anonymized) is given in the last subspell, this value is copied to the harmonized spell.

6 A reverse change from permanent to temporary within the same job is not considered very realistic.

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).[7] 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 sub-spells, 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 25 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 SC5
** version of this Scientific Use File
global version 18-0-0
** path where the data can be found on your local computer
global datapath Z:/Data/${cohort}/${version}
```

7 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).

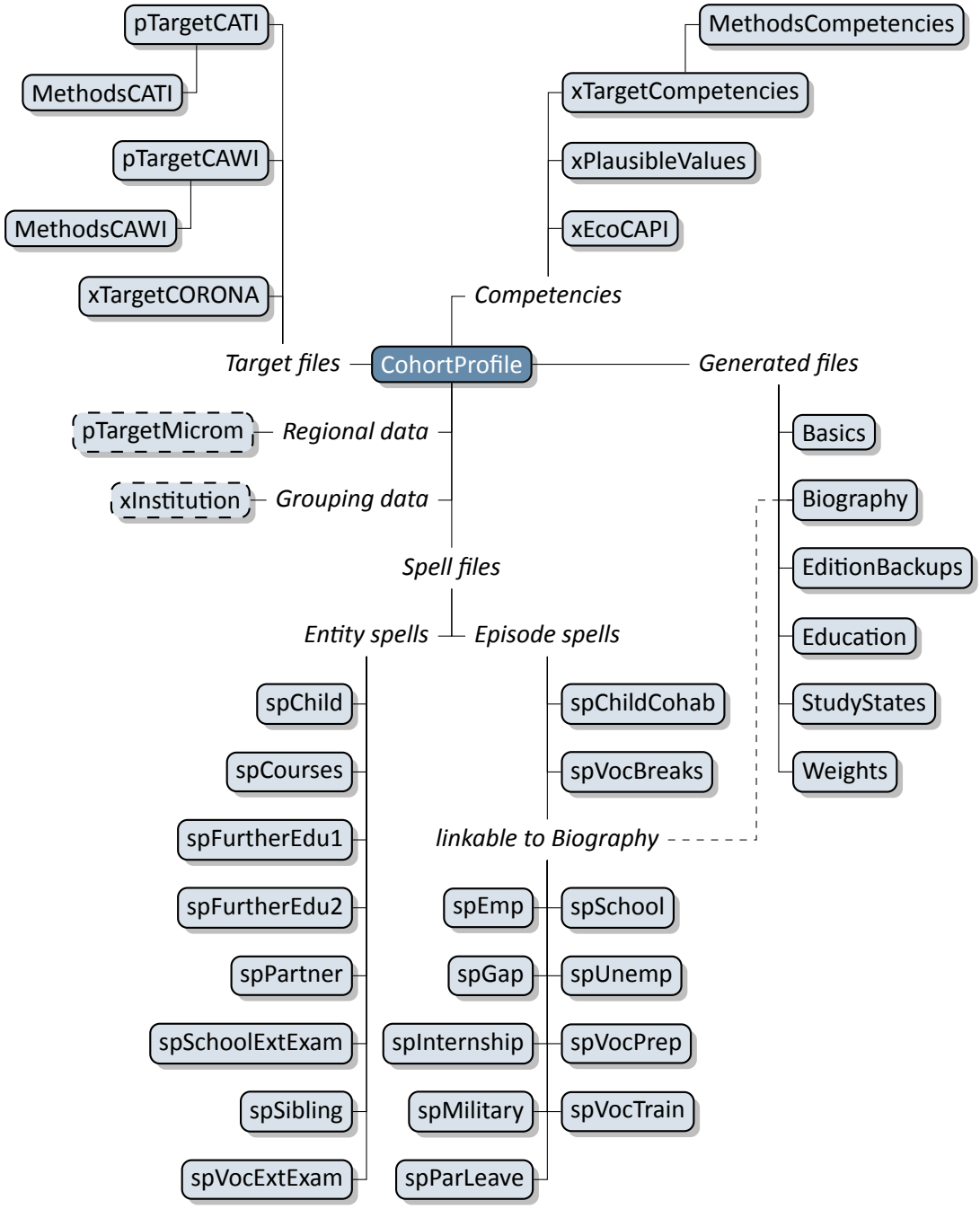

Figure 25: Graphical overview of all data files. Each node represents one data file. Relations are indicated by connection lines. Files with a dashed border are not available in the Download version of the Scientific Use File. Click on a data file to get more information.

4.5.1 Basics

« go back to overview

Description

Simplified information about respondents in a plain format

File structure

wide format: 1 row = 1 respondent

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

none

Number of variables / number of rows in file

84 / 36,360

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
tx29000	Age at interview month (years)
t70000m	Date of birth: month
t70000y	Date of birth: year
t700001	Gender
tx29003	Mother tongue: German
tx29004	Citizenship: German
tx29005	Born in Germany
t741001	Size of household (persons)
tx29060	currently employed
tx29904	Main spells of type 'Emp' (number)
tx29007	Age at migration to Germany

Exemplary data snapshot

ID_t	tx29000	t700001	tx29005	t741001	tx29060	tx29904
7007037	33.50	[m] male	yes	2	yes	5
7002271	30.75	[w] female	yes	2	yes	2
7004910	28.00	[m] male	no	2	yes	3
7010034	26.00	[m] male	yes	4	yes	4
7011790	27.25	[w] female	yes	1	yes	2

This file contains the latest reported basic information on each respondent, e. g., sociodemographic variables like age in years (`tx29000`), born in Germany (`tx29005`), gender (`t700001`), currently employed (`tx29060`), but also household characteristics, etc. It also contains meta information about some episodes like the number of main employment spells (`tx29904`). This data is generated from the `pTarget` files and a number of spell files. The Basics file is updated prospectively. That is, the file is cross-sectional (i. e., one row per person) and always includes updated information from the latest panel wave a respondent has participated. This simplified data structure can help to gain a first insight in the data. However, it should be handled with care, as it may not feature the *best* information about the respondent. This dataset only contains data from CATI interviews, information from CAWIs is not integrated. **Please use this file only to get a first overview of the data. Use the original panel or episode files for analyses!**

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

```
** open the data file
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear

** merge the data from Basics, enhancing every entry in CohortProfile
** (i.e. every wave, this is why m:1 merge is needed)
** with information from Basics
merge m:1 ID_t using ${datapath}/SC5_Basics_D_${version}.dta

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

** tabulate gender by wave
tab wave t700001

** please note that now, you have the most recent information known about respondents
** in every wave. This does not have to be equal to the information actually surveyed
** in that wave!
** Proceed at your own risk!
```

4.5.2 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 / 234,486

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
7003364	360002	3	36	2012	2012
7005482	240002	1	24	2010	2011
7012763	360001	5	36	2012	2012
7014322	260005	10	26	2015	2015
7018134	360005	5	36	2012	2012

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`, `spInternship`,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 2: 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
```

4.5.3 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_i ID_tg

Number of variables / number of rows in file

19 / 322,362

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
cohort	NEPS Starting Cohort
tx80220	Participation/drop-out status
tx80521	Data available: interview target person
tx80522	Data available: competence data target person
tx8610m	Competence testing Target person: survey month 1
tx8610y	Competence testing Target person: survey year 1
tx8600y	Survey Target person: survey year
tx8600m	Survey Target person: survey month
tx80524	Data available: institution
tx80107	Sample: first participation in wave

Exemplary data snapshot

ID_t	wave	tx80220	tx80521	tx80522	tx8610y	tx8600y	tx80524
7011366	1	Participation	yes	yes	2011	2011	yes
7011366	2	Temporary drop-out	no	missing by design	-54	-56	not determinable
7011366	3	Temporary drop-out	no	missing by design	-54	-56	not determinable
7011379	1	Participation	yes	yes	2011	2010	yes
7011379	2	Participation	yes	missing by design	-54	2011	yes
7011379	3	Participation	yes	missing by design	-54	2012	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 meta information like the ID of the institution (`ID_i`), various variables indicating participation (`tx80220`), availability of survey (`tx80521`), or availability of test data (`tx80522`). In addition, there are variables of the dates when the competence tests (`testm/y`) and the interview (`intm/y/d`) took place.

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

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

```
** open the data file
use ${datapath}/SC5_CohortProfile_D_${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
```

4.5.4 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 / 25,691

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
7006257	1	pTargetCATI	t731306	ID_t wave	5.00	2.00
7006525	1	pTargetCATI	t731306	ID_t wave	5.00	2.00
7012157	1	pTargetCATI	t731306	ID_t wave	5.00	2.00
7018354	10	spVocExtExam	ts15304	ID_t wave exam	28.00	19.00
7018914	1	pTargetCATI	t731306	ID_t wave	5.00	2.00

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 4: Working with EditionBackups

```
** In this example, we want to restore the original
** values in variable tg51410 (Intended degree) in 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=="pTargetCAWI" & varname=="tg51410"

** 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 sourcevalue_num editvalue_num

** rename the variables to emphasize affiliation
rename sourcevalue_num tg51410_source
rename editvalue_num tg51410_edit

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

** open pTargetCAWI
use ID_t wave tg51410 using ${datapath}/${cohort}_pTargetCAWI_D_${version}.dta, clear

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

** check all edition made
list ID_t wave tg51410* if _merge==3, nolab

** replace the variable in the datafile with its original value
replace tg51410=tg51410_source if _merge==3
```

4.5.5 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 splink

Other ID variables useful for linkage

tx28100

Number of variables / number of rows in file

12 / 60,971

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
7001974	1	2003	0	-20	0	220001	22
7001974	2	2007	3	10	2	220002	22
7001974	3	2010	5	13	3	220003	22
7001974	4	2015	8	18	9	240001	24
7001975	1	1999	0	-20	0	220001	22
7001975	2	2005	3	10	2	220002	22
7001975	3	2006	5	13	3	220003	22
7001975	4	2008	6	15	6	240001	24
7001975	5	2012	7	16	9	240002	24

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 5: Working with Education (find R example here)

```
** we want to merge the school type from spSchool to this datafile.
** For this to work, we first have to prepare spSchool and keep only
** harmonized episodes (subspell==0)
use ${datapath}/SC5_spSchool_D_${version}.dta, clear
label language en
keep if subspell==0
tempfile temp
save `temp'

** now, open the Education data file
use ${datapath}/SC5_Education_D_${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

** 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
```

4.5.6 MethodsCATI

« go back to overview

Description

Paradata from the targets CATI 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

48 / 145,654

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
ID_int	Interviewer: ID
wave	Wave
tx80200	Interview: number of all contact attempts
tx80207	Interview: response code differentiated (final outcome)
tx80400	Willingness: panel participation
tx80108	Type of recruitment
tx80301	Interviewer: gender
tx80302	Interviewer: age group
tx80209	Interview: length of interview (minutes)
tx80401	Willingness: Merging data from federal employment agency
tx80304	Interviewer: working experience as interviewer for infas

Exemplary data snapshot

ID_t	ID_int	wave	tx80200	tx80207	tx80301	tx80302	tx80209
7001968	1028	1	7	18	2	50-65 years	30.55
7001968	1405	3	3	50	2	up to 29 years	..00
7001969	1111	1	1	18	1	50-65 years	39.05
7001969	-54	3	52	33	.	.	..00

This dataset offers a variety of information on the data collection, e. g., gender (`tx80301`) and age (`tx80302`) of the interviewer; interview date (`intm`, `inty`); interview duration (`tx80209`); incentives (`tx80210`); and individual survey participation (`tx80220`).

Importantly, MethodsCATI contains all contacted respondents whether an interview was realized or not. Thus, MethodsCATI includes more cases than pTargetCATI.

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

```
** open the data file
use ${datapath}/SC5_MethodsCATI_D_${version}.dta, clear

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

** check out participation status by wave
tab wave tx80220

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

4.5.7 MethodsCAWI

« go back to overview

Description

Paradata from the targets CAWI 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

none

Number of variables / number of rows in file

21 / 29,915

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
tx80208	Interview: length of questionnaire (minutes)
tx80225	Interview: last delivery status
tx80250	Interview: winners of the lottery
tx80200	Interview: number of all contact attempts
tx80206	Interview: number of interruptions
tx80207	Interview: response code differentiated (final outcome)

Exemplary data snapshot

ID_t	wave	tx80208	tx80250	tx80206
7001982	11	14.11	0	0
7001982	14	-54.00	-54	-54
7001982	17	20.57	0	0
7002030	11	25.96	0	0
7002030	14	18.14	-54	0
7002030	17	15.39	0	0
7002115	11	26.18	0	0
7002115	14	21.80	-54	1
7002115	17	22.02	0	2
7002191	11	19.14	0	0

This dataset offers a variety of information on the data collection, e. g., teacher over-sample (`tx80122`); interview duration (`tx80208`); winners of the prize draw (`tx80250`); and the number of interruptions during the interview (`tx80206`).

Importantly, MethodsCAWI contains all contacted respondents whether an interview was realized or not. Thus, MethodsCAWI includes more cases than pTargetCAWI. MethodsCAWI provides data from wave 11 onwards because paradata on CAWI data prior to wave 11 was collected in a different way. Perhaps paradata for earlier waves will be included in future releases.

Stata 7: Working with MethodsCAWI

```
** open the data file
use ${datapath}/SC5_MethodsCAWI_D_${version}.dta, clear

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

** check out participation status by wave
tab wave tx80220

** how many waves have CAWI method data?
tab wave

** create one single variable containing the interview date
generate intdate=mdy(intm,intd,inty)
format intdate %td
list intd intm inty intdate in 1/10
```

4.5.8 MethodsCompetencies

« go back to overview

Description

Paradata from the targets competency tests

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_i ID_tg ID_int

Number of variables / number of rows in file

73 / 29,417

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
ID_i	Institution ID
wave	Wave
ID_tg	Test group ID
testm	Test: Survey day (month)
testy	Test: Survey day (year)
tx80422	Survey mode (realized)
ID_int	Interviewer: ID
tx80301	Interviewer: gender
tx80302	Interviewer: age group
tx80303	Interviewer: highest school-leaving qualification
tx80661	Number participants
tx80628	Questions about test tasks
tx80629	Questions about the calculator
tx80633	Signature

Exemplary data snapshot

ID_t	wave	ID_int	tx80301	tx80302	tx80303
7002115	12	-54	-54	missing by design	-54
7002163	1	1385	1	30-49 years	7
7002163	12	-54	-54	missing by design	-54
7002189	1	1385	1	30-49 years	7
7002189	12	-54	-54	missing by design	-54
7002204	1	1318	2	30-49 years	2
7002204	5	1466	2	50-65 years	18
7002204	12	-54	-54	missing by design	-54

Parallel to other Methods files, this dataset contains information about the testing situation, like durations, dates, interviewer IDs (`ID_int`), information about the interviewer (e. g., sex (`tx80301`), age (`tx80302`), and education (`tx80303`)), individual survey participation (`tx80220`), number of participants (`tx80661`), and disruptions and influences during testing (`tx80619`).

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

```
** open the data file
use ${datapath}/SC5_MethodsCompetencies_D_${version}.dta, clear

** how many respondents have been tested together in a group
bysort ID_tg: generate groupsize=_N if ID_tg>0 & !missing(ID_tg)
summarize groupsize

** create duration of math test; to achieve this, you first have to edit
** both start and end variables (which are stored in time format h:mm)

foreach var in tx80603 tx80604 {  // do the following for both variables
** convert to string, add leading zero
 tostring `var', gen(`var'_str) format(%04.0f)
** generate the etc datetime (ms. since 01jan1960 00:00:00.000)
** take care of missing values!
 gen `var'_ms=clock(`var'_str,"hm") if `var'>0 & !missing(`var')
}
** now the duration is the subtraction of start from end.
** this is recoded then from miliseconds to minutes
generate duration = (tx80604_ms - tx80603_ms)/(60*1000)

summarize duration
```

4.5.9 pTargetCATI

« go back to overview

Description

Data from respondents CATI questionnaires

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_i

Number of variables / number of rows in file

1,050 / 107,029

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
ID_i	Institution ID
wave	Wave
t431000	Migration sentiment
t531214	Tuition loan
t724403	Post-recording final grade
t531250	Source of finance: family
tg24503	Employment context doctorate
t712001	Kindergarten
t700001	Gender
t70000y	Date of birth: year
t514001	Satisfaction with life
t514008	Satisfaction with course of study
t741001	Size of household
t520003	Weight in kg

Exemplary data snapshot

ID_t	wave	t724403	tg24503	t700001	t70000y	t514008
7014830	13	4.0	4	[w] female	1991	7
7016246	12	..0	5	[m] male	1990	8
7016279	12	-93.0	4	[w] female	1991	8
7017128	12	1.4	4	[w] female	1991	9
7019256	13	1.1	3	[m] male	1992	7

The data in file `pTargetCATI` are from computer assisted telefone interviews (CATI). As many questions are asked repeatedly over different waves, data integration follows a long data format. This means, for each wave participated, there is an additional line for each participating target in this wave. Therefore, targets are uniquely identified by `ID_t` but lines are unique identified by `ID_t` and `wave` together. As there are only lines within `pTargetCATI` for persons who responded, there are less lines in `pTargetCATI` than in `CohortProfile`.[8]

This file contains hundreds of variables, which is the gross of all items surveyed. Some of them are sociodemographic like gender (`t700001`), year of birth (`t70000y`), country of birth (`t405010_g2`), or spoken languages (`t414000_g2`). Others are repeatedly administered in different waves (e. g., financial means for studying (`t531260`), satisfaction with studies (`t514008`)).

8 includes all students of the panel sample regardless of their questionnaire participation.

The file also includes information on the study program the respondents had started in winter term 2010/2011. The data were collected in the initial questionnaire and are stored in variable `tg01003_g1` (type of higher education institution), variables beginning with `tg0400` (different classifications of up to three subjects, information on majors or minors), variables `tg02001`, `tg02001_g1` (intended degree), variables `tg03001_g1`, `tg03001_g2` (type of intended teaching degree), and variables `tg15207_g1R`, `tg15207_g2R` (location of the higher education institution).

The initial questionnaire was mainly administered as a self-administered written survey. Partly, it was integrated into the first telephone interview. In this case, only the basic questions were asked. To avoid overstraining the participants, questions of minor relevance were omitted. Among these questions are those on admission restrictions (`tg10001`, `tg11001`, `tg11002`, `tg11003`), availability, use and quality of measures to facilitate integration into higher education (variables beginning with the string `tg0800`), and attitudes of parents and peers towards the study decision (variables beginning with `tg1500`).

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

```
** open the CohortProfile
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear

** merge some variable from pTargetCATI
merge 1:1 ID_t wave using ${datapath}/SC5_pTargetCATI_D_${version}.dta, ///
  keepusing(t400500_g1 t525204) nogen assert(master match)

** note that this information in now available only in waves which have
** surveyed the topic
tab wave t400500_g1

** if it makes sense, you can copy this information to cells of other waves.
** This copies information downwards (i.e., to later waves), unless a new
** value has been reported (which is usually what you want in a panel study)
bysort ID_t (wave): replace t400500_g1=t400500_g1[_n-1] ///
        if t400500_g1==-54 | missing(t400500_g1)

tab wave t400500_g1
```

4.5.10 pTargetCAWI

« go back to overview

Description

Data from respondents CAWI questionnaires

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_i

Number of variables / number of rows in file

1,530 / 59,749

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
ID_i	Institution ID
t242020	Quality equipment: literature
t242107	Higher education institution activities: sport
t289902	Living in shared living
t514001	Satisfaction with life
t272061	Motivation for courses/trainings
t30300b	Amount of rent
tg51004	Course of studies dropped out/interrupted/completed
tg74011	Time budget: work on doctorate
t241011	Time budget semester: classes
t241012	Time budget semester: self-study
tg72313	Discourse participation: lectures

Exemplary data snapshot

ID_t	wave	t289902	t514001	t30300b	tg51004
7003329	8	1	7	300	2
7005379	8	1	9	210	2
7008200	8	1	8	390	2
7013948	8	1	7	800	2
7015760	8	1	8	500	2

Apart from computer assisted telefone interviews (CATIs), data collection via computer assisted web interviews (CAWIs) has been conducted. `pTargetCAWI` also covers similar constructs collected in the CATI. There are items related to the amount of rent (`t30300b`), satisfaction with life (`t514001`), having a roommate (`t289902`), and there are also variables to help you to identify if a target is currently studying (`tg51000, tg51001, tg51004`). In contrast to CATIs, CAWIs are self-administered. Furthermore, biographical data such as episodes of employment or episode of vocational training were not collected.

Since wave 11, data on the device used during the online survey (`tg5910*`), the screen size (`tg5911*`), and also the survey setting (`tg5920*`) are collected and published in the SUF. These data enable new possibilities of method research. Please find more information about those variables via codebook, infoquery, or NEPSplorer (see section 1.2 and section 1.8).

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

```
** open pTargetCAWI
use ${datapath}/SC5_pTargetCAWI_D_${version}.dta, clear

** only keep a single variable, and IDs
keep ID_t wave t289902

** suppose you want to know if somebody ever lived with roommates.
** Then you could make use of the expression "t289902==1", which is true (1)
** if there has been a roommate, or false (0) otherwise. The maximum of
** this expression over waves results in 1 if any wave ever evaluated to true,
** and 0 otherwise.
egen roommate = max(t289902==1), by(ID_t)

** only keep this variable; as all waves contain the same information, we
** can fall back to cross-sectional structure
keep ID_t roommate
duplicates drop
tempfile room
save `room', replace

** finally, open CohortProfile and merge this variable
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear
merge m:1 ID_t using "`room'", nogen

tab wave roommate
```

4.5.11 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 / 197,552

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
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
7009879	7	1	145167	8	6	1
7009879	7	2	239686	7	.	2
7009879	7	3	305174	8	.	2
7009879	7	4	426799	7	.	.
7009879	7	5	503553	9	.	2

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 in wave 5 (data for waves 1, 3 and 7 have been enriched at a basic level).

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 11: 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
```

4.5.12 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

49 / 18,363

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

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

Exemplary data snapshot

ID_t	child	subspell	wave	ts3320y	ts33203	ts33204
7007074	1	1	10	2015	[w] female	biological child
7010725	1	1	15	2018	[w] female	biological child
7012075	1	1	16	2019	[m] male	biological child
7015398	1	1	16	2019	[m] male	biological child
7018615	1	1	1	1978	[m] male	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 12: Working with spChild (find R example here)

```
** open the data file
use ${datapath}/SC5_spChild_D_${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
```

4.5.13 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

20 / 5,087

Contains data from waves

1 2 **3** 4 **5** 6 **7** 8 **9** **10** 11 **12** **13** 14 **15** **16** 17 **18**

Exemplary variables

ID_t	ID target
child	Child number
spell	Spell number cohabitation with child
subspell	Number of subspell
wave	Wave
ts3331m	Start date living with child (month)
ts3331y	Start date living with child (year)
ts3332m	End date Living together with child (month)
ts3332y	End date Living together with child (year)
ts3332c	Current living together with child
ts33308	Episode update living together with child

Exemplary data snapshot

ID_t	child	spell	subspell	wave	ts3331y	ts3332y
7004432	2	203	2	5	2011	2013
7009331	2	202	2	5	2011	2013
7010971	1	101	2	5	2011	2013
7011894	1	101	1	3	2011	2012
7033771	1	101	2	5	2012	2013

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 13: Working with spChildCohab (find R example here)

```
** open the data file
use ${datapath}/SC5_spChildCohab_D_${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
** which gives you a cross-sectional display of cohabitation time for every
  respondent
keep ID_t total_duration_per_target
duplicates drop
```

4.5.14 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_w1 course_w2 course_w3

Number of variables / number of rows in file

49 / 10,892

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
splink	Link for spell merging
sptype	Spell type
t271001	Total duration of training courses
course_w1	Course number
t271011_w1	Course duration
course_w2	Course number
t271011_w2	Course duration
course_w3	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
7004760	13	260003	26	1301	1302	1303
7009996	15	260003	26	1501	1502	1503
7010681	13	260006	26	1301	1302	1303
7012312	18	260004	26	1801	1802	1803
7019308	12	260004	26	1201	1202	1203

This module comprises courses and trainings attended during episodes of employment (`spEmp`), unemployment (`spUnemp`), parental leave (`spParLeave`), military, or civilian service (`spMilitary`), as well as episodes from the `spGap` module. It comprises all spells from the past 12 months prior to the first interview that were recorded in the modules mentioned above. For follow-up interviews data on courses is collected up to three years retrospectively in case of temporary drop-outs between waves. The starting and end dates of the spells in this module represent the original starting and end dates of episodes (in which a course was taken) but not the start and end of the courses themselves.

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. For each of these episodes, information on up to three courses is included in wide format and therefore the course enumerators 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 datafile `Education`. If your interest in this data is not too profound, you are best advised to use `Education` instead.

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

```
** open the data file
use ${datapath}/SC5_spCourses_D_${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 ${datapath}/SC5_spEmp_D_${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
```

4.5.15 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

wave

Number of variables / number of rows in file

162 / 166,208

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
splink	Link for spell merging
subspell	Number of subspell
spell	Spell number
ts23222	In partial retirement, (active phase)
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
7012423	2	4	2013	2017	300	2
7012271	1	6	2016	2017	1180	3
7018396	1	3	2016	2018	1350	9
7007606	1	11	2014	2015	930	3
7014065	2	7	2015	2018	2650	9

This extensive module covers all spells of regular employment, including traineeships, preparatory service (e. g, for the teaching and legal profession), and internships (only in case that the target persons are not studying). Information on internships while studying is included in `spInternship`. 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 professional position (`ts23203`), net income (`ts23410`), relevance to degree course (`tg26190`), or permanent contract (`ts23320`), type of student employment (`tg2608b`), quality of student jobs (`t265401-t265423`) and internships (`tg26300-tg2630i`). Have a look at `pTargetCATI` and `pTargetCAWI` for more fine-grained information on teacher training and the situation of teachers.

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

```
** open the data file
use ${datapath}/SC5_spEmp_D_${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 ${datapath}/SC5_Biography_D_${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
```

4.5.16 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

18 / 9,320

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
course	Course number
t271048	Course is ongoing
t271049	Termination course
t272000_O	Content other course
t271050	Other courses 2
t271051	Other course
t272000_g13	Content other course (course ID)
t271043	Duration of course (hours)
tx80211	Survey/Test instrument

Exemplary data snapshot

ID_t	wave	course	t271048	t271050	t271051
7006610	18	1801	no	no	no
7011719	12	1205	no	no	no
7012620	9	902	no	no	no
7015025	13	1301	yes	no	no
7015178	9	901	no	no	no

This module contains information on further courses (also private courses) since the last interview 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).

Stata 16: Working with spFurtherEdu1 (find R example here)

```
** open the datafile
use ${datapath}/SC5_spFurtherEdu1_D_${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 CohortProfile). So first, we have to remodel
** the file so one row contains one wave. We do this by Stata's reshape command
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 ${datapath}/SC5_CohortProfile_D_${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
```

4.5.17 spFurtherEdu2

« go back to overview

Description

information about 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

28 / 18,864

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
course	Course number
t279040	Professional/private reasons
t279046	Course costs employer
t272040	Provider
t279041	Motivation for course attendance
t272043	Certificate
t272003	Course evaluation: learned new things
t274022	Course evaluation: instructor patient
t279047	Course costs employment agency

Exemplary data snapshot

ID_t	wave	course	t279046	t279041	t272043
7010079	12	1203	not at all	a lot of effort	1
7010448	15	1501	not at all	some effort	1
7012961	15	1502	fully	a lot of effort	1
7017457	13	1301	fully	little effort	1
7019177	10	1001	fully	a lot of effort	1

The survey instrument randomly selected two courses from the `spCourses` and `spFurtherEdu1` modules, collecting additional information on these courses (e. g., costs incurred by employer `t279046`, motivation `t279041`, and certificates `t272043`). These data are included in `spFurtherEdu2`.

Stata 17: Working with spFurtherEdu2 (find R example here)

```
** Two possibilities to use spFurtherEdu2

** A) Merge data to spCourses

** open spCourses datafile
use ${datapath}/SC5_spCourses_D_${version}.dta, clear

** one row contains information for up to three courses.
** To make merging possible, you first have to reshape the datafile
** so one row contains only one course
reshape long course_w, i(ID_t wave splink) j(course_nr)
rename course_w course

** merge spFurtherEdu2 using ID_t and course
merge m:1 ID_t course using ${datapath}/SC5_spFurtherEdu2_D_${version}.dta, keep(
  master match)

** ----
** B) merge to spFurtherEdu1

** open spFurtherEdu1 datafile
use "${datapath}/SC5_spFurtherEdu1_D_${version}.dta", clear

** merge spFurtherEdu2 using ID_t and course
merge 1:1 ID_t course using ${datapath}/SC5_spFurtherEdu2_D_${version}.dta, keep(
  master match)
```

4.5.18 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

22 / 15,918

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
splink	Link for spell merging
spell	Spell number
subspell	Number of subspell
wave	Wave
spms	Check module: spell type
ts29901	Auxiliary variable current gap episode
ts29300	Episode mode
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	Training course during gap
ts29101	Type of gap episode

Exemplary data snapshot

ID_t	spell	subspell	wave	ts29901	ts2911y	ts2912y
7010046	2	1	1	1	2011	2011
7014580	2	1	1	1	2011	2011
7016871	1	0	1	1	2010	2010
7017115	1	1	1	1	2011	2011
7033526	2	1	1	1	2010	2012

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). The most common gap episode is (extended) holidays.

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

```
** open the data file
use ${datapath}/SC5_spGap_D_${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 ${datapath}/SC5_Biography_D_${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
```

4.5.19 spInternship

« go back to overview

Description

reported internship episodes

File structure

spell format: 1 row = 1 internship 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

40 / 39,435

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
splink	Link for spell merging
spell	Spell number
subspell	Number of subspell
wave	Wave
tg3607m	Start month internship episode
tg3607y	Start year internship episode
tg3608m	End month internship episode
tg3608y	End year internship episode
tg36109	Continuing duration of internship period
tg36110	Type of internship
tg36111	Average working hours Internship
tg36119	Placement as an intern
t265321	Learning content: autonomy 1
t264300	Support: Supervision

Exemplary data snapshot

ID_t	spell	subspell	wave	tg3607y	tg3608y	tg36111
7002759	2	2	5	2012	2012	7
7005673	4	2	10	2015	2015	35
7012943	1	2	7	2011	2011	42
7015812	3	2	9	2014	2014	30
7016577	3	2	3	2011	2011	40

As internships during studies are regarded as central to professional success, both compulsory and voluntary internships have been surveyed and made available in this datafile. Information about duration, renumeration, learning content, and other key aspects have been surveyed.

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

```
** open the data file
use ${datapath}/SC5_spInternship_D_${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 ${datapath}/SC5_Biography_D_${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
```

4.5.20 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

21 / 5,160

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
7007835	250001	1	1	7	2013	2014
7009797	250001	2	1	3	2008	2012
7010285	250001	2	1	3	2009	2011
7011558	250002	1	2	5	2012	2013
7017803	250001	5	1	9	2010	2015

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

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

```
** open the data file
use ${datapath}/SC5_spMilitary_D_${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 ${datapath}/SC5_Biography_D_${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 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

29 / 5,389

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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 (year)
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
7003701	2	202	2	18	2019	2020
7004385	1	101	2	18	2020	2021
7013591	1	101	1	7	2013	2014
7013937	1	101	1	13	2018	2018
7014305	1	101	2	18	2019	2021

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. The employment spell is not neccessarily interrupted if the mother is on parental leave as part-time work is legal.

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

```
** open the data file
use ${datapath}/SC5_spParLeave_D_${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 ${datapath}/SC5_Biography_D_${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
```

4.5.22 spPartner

« go back to overview

Description

history of partners

File structure

entity format: 1 row = 1 partner of 1 respondent

ID variables needed to identify a single row

ID_t partner subspell

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

109 / 84,415

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
partner	Partner number
subspell	Number of subspell
ts31204	Partner: born Germany/abroad
ts31211	Partner German
ts31203	Gender of partner
ts3141m	Marriage date (month)
ts3141y	Marriage date (year)
ts3120y	Year of birth partner
tg2811m	Start date partnership - month
tg2811y	Start date partnership - year
tg2804m	End date partnership episode (month)
tg2804y	End date partnership episode (year)
ts31206	Age at immigration Partner
ts31207	Place of birth Father Partner
ts31209	Place of birth Mother Partner

Exemplary data snapshot

ID_t	partner	subspell	ts31203	ts3120y	tg2811m	tg2811y	tg2804m	tg2804y
7004882	2	1	[m] male	1979	11	2005	8	2011
7014891	2	1	[m] male	1960	5	2005	5	2011
7013068	1	1	[m] male	1968	7	1994	6	2011
7004182	1	0	[m] male	1986	5	2011	10	2012
7014594	2	0	[m] male	1990	12	2013	4	2014

This module covers the partnership history of the respondent. Respondents' subjective reports define whether they live in a relationship and whether they cohabit or not. A comprehensive set of additional questions refers to partners since the beginning of winter term 2010. For earlier partners, only information on the year of birth and education is available. The enumerator variable partner identifies partners *within* respondents. This variable is coded 1 for the first partner and counts upwards until the last (current) partner.

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

```
** open the data file
use ${datapath}/SC5_spPartner_D_${version}.dta, clear

** switch to english language
label language en

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

** to find out if a respondent is or was ever been married,
** check out if the indicating variable ever stated a marriage
bysort ID_t: egen married = max(ts31410==1)

** look at the data
list ID_t partner ts31410 married in 1/20, sepby(ID_t)

** reduce the datafile, so you have one single row for each respondent
keep ID_t married
duplicates drop

** you now can save this datafile and merge it to, e.g., CohortProfile
tempfile married
save `married'
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear
merge m:1 ID_t using `married', nogen keep(master match)
```

4.5.23 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 / 47,058

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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	Starting month school episode
ts1111y	Starting year school period
ts1112m	Final month of school episode
ts1112y	Final year of school episode
ts11209	School-leaving qualification
ts11214	Intended school-leaving qualification
ts11218	Final grade school-leaving certificate
t724801	1st Abitur subject
t724802	2nd Abitur subject

Exemplary data snapshot

ID_t	splink	subspell	spell	wave	ts1111y	ts1112y
7006174	220001	0	1	1	1997	2001
7012883	220003	1	3	1	2008	2011
7016842	220001	0	1	1	1987	1991
7019499	220004	1	4	3	2011	2012
7019499	220004	2	4	5	2011	2012

This module covers each respondent's general education history from school entry until the date of completion, or in case of enduring episodes, date of interview. This includes

- 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 23: Working with spSchool (find R example here)

```
** open the data file
use ${datapath}/SC5_spSchool_D_${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 ${datapath}/SC5_Biography_D_${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 spSchoolExtExam

« go back to overview

Description

school certificates acquired by recognition or external students' examination

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

28 / 812

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
ts11301_g2	Country of awarded school-leaving qualification (categorized)

Exemplary data snapshot

ID_t	wave	exam	ts11300	ts1130y	ts11302	ts11300_g1
7011945	1	1	1	2010	.	1
7012231	1	1	1	2004	.	1
7017591	5	1	1	2012	5	1
7018475	7	1	1	2013	5	1
7018604	1	1	1	2007	.	1

`spSchoolExtExam` comprises information about school exam certifications that have not been acquired through "regular" schooling. 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 classes)
- certificates that are automatically awarded by advancing through grades in upper secondary education or by completing vocational training

Stata 24: 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 ${datapath}/SC5_pTargetCATI_D_${version}.dta, clear
keep if wave==1  // only first wave as this data is time-invariant
keep ID_t t70000m t70000y
label language en
tempfile temp
save `temp'

** now, open the data file
use ${datapath}/SC5_spSchoolExtExam_D_${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(t70000y,t70000m)
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)
```

4.5.25 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

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

11 / 26,932

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
sibling	Sibling number
tx80211	Survey/Test instrument
tg3270m	Month of birth sibling
tg3270y	Year of birth sibling
tg32706	Is sibling still alive?
tg32708	Employment status Sibling
tg32709	Unemployment siblings
tg32711	Highest school-leaving qualification siblings
tg32724	Sibling lives with parents

Exemplary data snapshot

ID_t	wave	sibling	tg3270m	tg3270y	tg32708	tg32711
7002452	1	1	12	1987	unemployed	5
7002996	1	2	11	1989	unemployed	5
7004068	1	2	1	1988	in a side job	5
7006622	1	1	12	1962	full-time employed	5
7011113	1	2	9	1981	unemployed	3

`spSibling` contains all siblings of the respondent reported in wave 1. Each sibling is stored in one row, containing information about the date of birth (`tg3270m/y`), employment status (`tg32708`), and highest degree (`tg32711`).

Stata 25: 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 ${datapath}/SC5_pTargetCATI_D_${version}.dta, clear
keep if wave==1  // keep wave-1-data only as data on sibling was only collected once
keep ID_t t70000m t70000y
label language en
tempfile temp
save `temp'

** now, open the spSibling data file
use ${datapath}/SC5_spSibling_D_${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(tg3270y,tg3270m)
gen  target_bdate=ym(t70000y,t70000m)
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
```

4.5.26 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

28 / 20,455

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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
ts2512y	End unemployment episode
ts25202	Receipt of unemployment benefits or support at the beginning
ts25203	Registered unemployment currently the case/finished
ts25205	Number job applications
ts25206	Invitation to job interviews
ts25207	Number job interviews

Exemplary data snapshot

ID_t	subspell	spell	wave	ts2511m	ts2511y	ts2512m	ts2512y
7002459	1	1	7	11	2013	7	2014
7003127	2	3	13	9	2017	12	2017
7007302	1	1	9	10	2014	6	2015
7011032	2	6	7	2	2013	6	2013
7018815	1	3	10	9	2015	5	2016

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 26: Working with spUnemp (find R example here)

```
** open the data file
use ${datapath}/SC5_spUnemp_D_${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 ${datapath}/SC5_Biography_D_${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 spVocBreaks

« go back to overview

Description

breaks during vocational training

File structure

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

ID variables needed to identify a single row

ID_t splink break

Other ID variables useful for linkage

none

Number of variables / number of rows in file

17 / 2,063

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
splink	Link for spell merging
break	Training interruption number
ts1531y	Start year Interruption of training episode
ts1531m	Start month Interruption of training episode
ts1532y	End year interruption of training episode
ts1532m	End month interruption of training episode
tg2419a	Status during interruption: semester off
tg2419b	Status during interruption: de-registered
tg2419c	Status during interruption: not formally de-registered

Exemplary data snapshot

ID_t	splink	break	ts1531y	ts1532y
7002591	240001	1	2012	2012
7016076	240003	1	2013	2014
7005766	240002	1	2013	2014
7002818	240001	1	2011	2012
7006310	240001	1	2012	2013

This module covers all breaks of further trainings, vocational and/or academic, that a respondent ever attended – with a special focus on academic education. Information on vocational breaks were part of `spVocTrain` in prior data releases. Since release 16-0-0 break episodes are being extracted and edited to `spVocBreaks`. The data structure of breaks has been transformed from wide format to long format. In this dataset …

- different types of breaks (break semesters, de-registrations, non-formal breaks) are included.
- includes several breaks within a single episode of a person.
- closes gaps between succeeding breaks (< 3 months) and combines overlapping breaks to continuous breaks.
- breaks within breaks were deleted.
- dates of beginnings and endings were corrected and stored as variables with `_g1`–suffixes.

- every break is in a separate row.
- `splink` helps you to merge data to `spVocTrain` and `Biography` as well.

Stata 27: Working with spVocBreaks

```
** example 1: merge spVocBreaks and spVocTrain
** open the vocational breaks
use ${datapath}/SC5_spVocBreaks_D_${version}.dta , clear

** reshape study breaks to wide format to match data with spVocTrain; first add _w-
  suffix to variables
foreach var of varlist ts15310_g1 ts15310_g2 ts1531y ts1531m ts1531m_g1 ts1531y_g1
  ts1532c ts1532y ts1532m ts1532y_g1 ts1532m_g1 tg2419a tg2419b tg2419c {
        rename `var' `var'_w
}
reshape wide *_w, i(ID_t splink) j(break)

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

** open the data file
use ${datapath}/SC5_spVocTrain_D_${version}.dta, clear

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

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

** you now have put together information of breaks and the vocational track to
  analyze the students with breaks. The number of total episodes with breaks reduces
  the amount of rows of the combined dataset.


** example 2: merge spVocBreaks and Biography (further data preparation to analyze
  data is recommended)
** open the vocational breaks
use ${datapath}/SC5_spVocBreaks_D_${version}.dta , clear , clear

*merge breaks with biography data
merge m:1 ID_t splink using ${datapath}/SC5_Biography_D_${version}.dta, clear

** now you could cut those vocational episodes using dates of episodes and breaks to
  re-define vocational episodes
************************
```

4.5.28 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

30 / 3,351

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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 external examination (month)
ts1530y	Date external examination (year)
ts15304	External examination qualification
ts15302	External examination in Germany/abroad
th28370	External examination preparation done abroad for at least one month

Exemplary data snapshot

ID_t	wave	exam	ts1530m	ts1530y	ts15304
7003689	12	1	10	2016	30
7007157	16	2	7	2019	30
7011162	15	1	6	2019	30
7019029	13	1	7	2017	30
7020496	15	1	10	2018	30

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 28: 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 ${datapath}/SC5_pTargetCATI_D_${version}.dta, clear
keep if wave==1  // only first wave as this data is time-invariant
keep ID_t t70000m t70000y
label language en
tempfile temp
save `temp'

** now, open the data file
use ${datapath}/SC5_spVocExtExam_D_${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(t70000y,t70000m)
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)
```

4.5.29 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

22 / 504

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

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

Exemplary data snapshot

ID_t	spell	subspell	wave	ts1311m	ts1311y	ts1312m	ts1312y
7004383	1	2	13	10	2016	9	2017
7008524	2	2	16	5	2019	6	2019
7009208	1	2	10	9	2014	8	2015
7010862	1	1	5	8	2012	3	2013
7023084	1	1	10	1	2016	4	2016

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 29: Working with spVocPrep (find R example here)

```
** open the data file
use ${datapath}/SC5_spVocPrep_D_${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 ${datapath}/SC5_Biography_D_${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
```

4.5.30 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

165 / 135,766

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

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

Exemplary data snapshot

ID_t	spell	subspell	ts1511m	ts1511y	ts1512m	ts1512y
7006893	2	3	4	2011	4	2014
7013538	1	3	10	2010	6	2013
7015953	1	2	10	2010	6	2012
7022112	3	1	8	2014	7	2015
7015566	3	2	10	2013	7	2015

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

- tertiary education at universities (including colleges of education, theology, and art and music), universities of applied sciences, Berufsakademien/cooperative state universities, colleges of public administration). Note: Up to three subjects (majors and minors) are recorded.
- doctoral or postdoctoral studies
- 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

In case of higher education study, new episodes are generated if

- a subject changes over the course of studies, or
- the intended degree changes over the course of studies (e. g., from master's degree to state examination), or
- the higher education institution changes.

If a higher education episode follows immediately a preceding higher education episode, interviewees are asked whether the type of degree was changed (`tg24146`), whether different subjects were chosen (`tg24159`) or whether the respondent moved to another higher education institution (`tg24121`). New information on the intended degree (`ts15221`), subjects (variables beginning with `tg2416`, `tg2417`), and higher education institution were only collected when the aforementioned questions were answered with *yes*. In case of a negative answer, the variable `ts15221` (intended degree) takes the value of the preceding episode while the variables containing information on the subjects take the value -29 (value from the last sub-episode). The information of the preceding episode was integrated into the service variables `tg24162_g1`, `tg24165_g1`, `tg24168_g1`, `tg24170_g1`–`tg24170_g5`, `tg24173_g1`–`tg24173_g5` and `tg24176_g1`–`tg24176_g5` (see section section 5.1.1).

Information on the subjects, intended degree, and the higher education institution of the first study episode in winter term 2010/2011 was collected in the initial questionnaire, which was mainly administered as a written survey and partly integrated into the first telephone interview. This data can be found in `pTargetCATI` in the variables `tg0400*` (information on the subjects), `tg01003_g1` (type of higher education institution), `tg15207_g1R`, `tg15207_g2R` (location of the higher education institution), and `tg02001*` (intended degree). The information was not newly collected in the first telephone interview but was integrated into the service variables `tg2417*` and `tg01003_ha` (see section section 5.1.1). The variable `h_aktstu` in `spVocTrain` indicates which episode refers to the first study episode in winter term 2010/2011 (`h_aktstu==1` "Episode is 1st study episode WT 2010 (start of study)").

In the telephone interview following wave 1 (i. e., in wave 3, 5 or 7, depending on panel participation), the vocational education history from winter term 2010/2011 onwards was newly collected with an improved survey instrument. This has led to duplicate and/or right-censored episodes in the dataset `spVocTrain`. In order to deal with those episodes, the variable `tx20100` was introduced to give a recommendation which episodes should be used for analyses. The rule applies that episodes from wave 1 are always recommended when the start date lies at or before the beginning of the first study episode of the winter term 2010/11. Episodes from wave 1 are never recommended when the start date lies after the beginning of the first study episode of the winter term 2010/11.

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

```
** open the data file
use ${datapath}/SC5_spVocTrain_D_${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 ${datapath}/SC5_Biography_D_${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 StudyStates

« go back to overview

Description

Data on state of studies derived from spVoc-Train

File structure

long format: 1 row = 1 respondent in 1 wave

ID variables needed to identify a single row

ID_t wave or ID_t tx24001

Other ID variables useful for linkage

ID_t tx24022

Number of variables / number of rows in file

51 / 206,783

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
tx24000	Recommendation: use person (less than 3 episodes per wave)
tx24001	Chronological order of the times of the interviews
tx24100	Status of studies (completed,ongoing)
tx24101	Status of the course of study (CAWI)
tx15318	Successful completion of the training
tx24011	Change of higher education institution: change from higher education institution
tx92401	Type of higher education institution: previous higher education institution
tx92402	Type of higher education institution: current higher education institution

Exemplary data snapshot

ID_t	wave	tx24000	tx24100	tx24101
7003370	4	yes	episodes ongoing, none completed	is currently studying
7009991	4	yes	episodes ongoing, none completed	is currently studying
7014381	11	yes	episodes ongoing, none completed	is currently studying
7016506	4	yes	episodes ongoing, none completed	is currently studying
7018025	2	yes	episodes ongoing, none completed	is currently studying

The file `StudyStates` contains all target persons for each wave as long as they have not dropped out. As soon as a person drops out, it will not be part of the dataset since that certain wave.

For each respondent in each wave, `StudyStates` contains information on status of studies/tertial education (`tx24100`), which vocational qualification is achieved (`tx15317`), which subjects were choosen or switched. There is also data on wether a person continued its studies at a different educational institution (`tx24011`) as well as information on study-breaks (`tx15310`, `tx24190`).

This dataset boils down information from spVocTrain to meet the data structure of long-format files such as pTargetCATI – one line for each person per wave. This procedure inevitable goes along with information loss. As some target persons' data show multiple educational episodes at the same time, defining a proper status of studies, for instance, is virtually impossible to achieve without sound assumptions. Therefore we introduced an indicator that recommends the usage of a person within the dataset (tx24000). The data is recommended to use as long as the person has less than three episodes in any wave.

This dataset is still a kind of beta version of the desired outcome, therefore we suggest to use this dataset to get a quick insight on data concering study states. Data will be polished and will be more usuable in the next release!

Stata 31: Working with StudyStates

```
*** 1. enriching StudyStates with episode data from spVocTrain ***

** open spVocTrain file
use "${datapath}/SC5_spVocTrain_D_${version}.dta" , clear

** only keep full or harmonized episodes and save file temporarily
keep if subspell == 0
tempfile spvoc
save "`spvoc'", replace

** open StudyStates file
use "${datapath}/SC5_StudyStates_D_${version}.dta" , clear

** rename tx24022 to splink and keep only valid episodes
rename tx24022 splink
keep if splink > 0

** merging StudyStates with spVocTrain, only keeping desired variables
merge m:1 ID_t splink using "`spvoc'", keep(master matched) nogenerate keepusing(
  tg24203 tg24205 tg24162_g1 tg24165_g1 tg24168_g1)


*** 2. merging StudyStates with pTarget-data ***

** open StudyStates file
use "${datapath}/SC5_StudyStates_D_${version}.dta" , clear

** merging StudyStates with pTargetCATI, only keeping desired variables
merge 1:1 ID_t wave using "${datapath}/SC5_pTargetCATI_D_${version}.dta", keep(master
   matched) nogenerate keepusing(t406000 t731351_g1 t31300a t34006k tg2411a
  t66003a_g1 t531260)

** merging data with pTargetCAWI, only keeping desired variables
merge 1:1 ID_t wave using "${datapath}/SC5_pTargetCAWI_D_${version}", keep(master
  matched) nogenerate keepusing(t291501 t291502 t321104 t66007a t513051 tg54112
  t30300b)
```

4.5.32 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

31 / 17,909

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
ID_i	Institution ID of sampling
sLevel_R	Stratification second level
ID_cl	Cluster ID
w_h	Weight stratification first level
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
w_allWaves	Longitudinal weight for all waves
w_allCATI	Longitudinal weight for all CATI waves
w_allCAWI	Longitudinal weight for all CAWI waves

Exemplary data snapshot

ID_t	ID_i	ID_cl	w_h	w_t1	w_t2	w_allWaves
7011465	1002288	300	6.36600	1.28003	1.03312	0.90726
7009761	1002997	31	6.28600	0.99914	1.02201	0.91164
7015569	1002134	355	6.36600	2.29541	1.87839	2.73801
7011231	1002017	137	6.28600	0.85650	0.68261	0.49203
7004434	1002103	97	6.28600	1.41769	1.34358	1.55826

Weighting variables (starting with `w_`) are included in the Weights dataset. Also, you find cluster (`ID_cl`) and stratification (`stratum`) identifiers here. `ID_i` resembles university at sampling, which took part prior to wave 1. 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 Zinn et al., 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 32: Working with Weights (find R example here)

```
** open Weights datafile
use ${datapath}/SC5_Weights_D_${version}.dta, clear

** note that this file is cross-sectional, although the weights
** seem to contain panel logic
d w_t*

** only keep weight corresponding to all waves
keep ID_t w_t123456789

** create a "panel" logic, i.e., clone each row
expand 9

** then create a wave variable
bysort ID_t: gen wave=_n

** save as temporary file
tempfile weights
save `weights', replace

** open CohortProfile
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear

** and merge weight
merge 1:1 ID_t wave using `weights', nogen

** note that this weight is only non-zero if respondents participated in
** all waves
tab wave tx80220 if w_t123456789!=0
```

4.5.33 xEcoCAPI

« go back to overview

Description

additional competencies for students of economics and business administration

File structure

wide format: 1 row = 1 student

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

wave ID_int

Number of variables / number of rows in file

136 / 600

Contains data from waves

1 2 3 4 5 6 **7** 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
tx80921	Participation status: Economics-subsample
testm	Test: Survey day (month)
testy	Test: Survey day (year)
bas7mar1_c	Economic competence: marketing 1
bas7_sc1	Economic competence: WLE
bas7_sc2	Economic competence: SE(WLE)
ID_int	Interviewer: ID
tg90308	Number semesters economics
tg24160_g2	Subject group subject 1 (destatis 2010/11)
tx80200	Interview: number of all contact attempts
tx80302	Interviewer: age group
tx80430	Interview: location

Exemplary data snapshot

ID_t	wave	tx80921	bas7_sc1	bas7_sc2	tg90308	tg24160_g2	tx80200
7004844	7	6	0.29365	0.40530	7	3	3
7005848	7	6	0.26627	0.40280	7	3	3
7009568	7	6	0.80859	0.41941	7	3	2
7012992	7	6	0.05863	0.37350	7	3	5
7014686	7	6	0.33918	0.38575	7	3	2

Apart from the basic CATI-data collection in wave 7, additional data was collected for students of economics and business administration. A paper-based competency test containing questions specificially for the target's field of study was embedded within a short computer assisted personal interview (CAPI).

This data was part of `pTargetCATI` and `xTargetCompetencies` in releases prior to data version 10-0-0. To emphasize the focus on this small subgroup of targets, all this information is now gathered in `xEcoCAPI`. As this file contains data from wave 7 only, `ID_t` is a unique identifier in this wide-format dataset. To make things simpler, participation in CAPI, CATI, and competency testing is indicated by `tx80921`. Additional methods data – like number of contact tries (`tx80200`) and reasons for item-nonresponse in testing (e. g., `tx80411`) – are availble as well.

CAPI data are basically focussing on the student's area of studies (e. g., `tg24160_g2`). For more information see Lauterbach (2015).

Stata 33: Working with xEcoCAPI

```
** open the CohortProfile
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear

** merge some variables from xEcoCAPI
merge 1:1 ID_t wave using ${datapath}/SC5_xEcoCAPI_D_${version}.dta, ///
  keepusing(bas7_sc1 bas7_sc2) nogen assert(master match)

** note that this information in now available only in waves which have
** surveyed the topic
tab wave bas7_sc1
```

4.5.34 xInstitution

« go back to overview

Description

context information about the institution

File structure

wide format: 1 row = 1 area of studies in 1 institution

ID variables needed to identify a single row

ID_i tg04001_g7

Other ID variables useful for linkage

none

Number of variables / number of rows in file

127 / 3,520

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_i	Institution ID
tg04001_g7	Subject group WT 2010 (for merging with context data)
tg91102_R	HEI region: BIK-region type
tg92104_O	HEI: Winner Cluster of excellence 2006 or 2007
tg92301_O	HEI: Funding body
tg92601_R	HEI: Students 2010 total (aggr. Tercent. universities)
tg93204_O	SG: Students 2010: male
tg93205_O	SG: Students 2010: female
tg93601_O	SG: Students per professor
tg93602_R	SG: Students per lecturer (aggr. by terc. of all HEI, sep. U/UAS)

Exemplary data snapshot

ID_i	tg04001_g7	tg92104_O	tg92301_O	tg92601_R
1002090	6	1	1	3
1002014	6	1	1	3
1002057	9	1	1	3
1002086	3	1	1	3
1002064	9	1	1	3

Data file `xInstitution` contains context data (e.g., size of the institution, regional unemployment rate) for all 413 higher education institutions which were listed in the codebook of the Federal Statistical Office in 2010/2011. However, higher education institutions with different locations are only considered once with combined information for all locations (for detailed information, see Weber, 2014).

Note that due to data protection issues, this file is not available in the Download version of SUF. You find it in **RemoteNEPS** and **Onsite**.

Please also note that the context information up to now has not been updated and refers to most recent information available in 2010.

Stata 34: Working with xInstitution (find R example here)

```
** open datafile
use ${datapath}/SC5_pTargetCATI_O_${version}.dta, clear

foreach var in ID_i tg04001_g7 { // do the following for both variables
** copy the information from the first wave downwards for each target,
** unless a new value has been reported
bysort ID_t: replace `var' = `var'[_n-1] ///
        if `var' == -54|missing(`var')
}
** drop all observations where no satisfaction with studies was reported
drop if t514008 == -98|t514008 == -97|t514008 == -93|t514008 == -54|missing(t514008)

** some respondents reported satisfaction with studies in 7th and in 9th waves
** to keep the latest information, create a seq and a max variables
bysort ID_t: gen seq = _n
bysort ID_t: gen max = _N
** only keep the latest reported iformation
keep if seq == max
** only keep the variables relevant for the merge and the analysis
keep ID_t ID_i tg04001_g7 t514008

** merge two variables from xInstitution
merge m:1 ID_i tg04001_g7 using ${datapath}/SC5_xInstitution_O_${version}.dta, ///
        keepusing(tg92601_R tg92104_O) nogen assert(master match)

** assuming that the less students at university the more intensive the support by
  the
** university staff per student and the more satisfied are students with their
  studies
** tabulate Satisfaction with studies by Students 2010 total
** note that the following analysis is feasible in both, RemoteNEPS and Onsite
tab t514008 tg92601_R, col

** assuming that students at excellence universities are more satisfied with
** their studies, tabulate the distribution of satisfaction by tg92104_O
** note that the following analysis is only feasible in the Onsite version of SUF,
** since the variable tg92104_O is anonymized in RemoteNEPS
tab t514008 tg92104_O, col
```

4.5.35 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

114 / 11,739

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2010/2011 (CATI+competencies))
wave_w5	Row contains data from wave 5
wave_w12	Row contains data from wave 12
mas1_pv1	Math: cross-sectional plausible value 1
mas1_pv2	Math: cross-sectional plausible value 2
mas1_pv10	Math: cross-sectional plausible value 10
mas1_pv1u	Math: longitudinal plausible value 1
mas1_pv2u	Math: longitudinal plausible value 2
mas1_pv10u	Math: longitudinal plausible value 10

Exemplary data snapshot

ID_t	wave_w1	mas1_pv1	mas1_pv2	mas1_pv10	mas1_pv1u
7015430	1	0.67231	1.13139	0.74893	1.54801
7008430	1	0.80125	0.44029	1.00561	0.14055
7014148	1	0.41126	0.39759	0.43051	1.71433
7012790	1	0.23946	0.86760	0.73805	1.54801
7014800	1	1.34065	2.87229	1.21171	2.71423

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 35: 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!
```

4.5.36 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

371 / 11,810

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2010/2011 (CATI+competencies))
wave_w5	Row contains data from wave 5
mas1r092_c	Mathematical competence: Item 2
mas1_sc1	Mathematical competence: WLE
mas1_sc2	Mathematical competence: SE(WLE)
res1_sc1	Reading competence: WLE
res1_sc2	Reading competence: SE(WLE)
rsci0051_c	Reading speed: Item 51
rss1_sc3	Reading speed: sum
ics3_sc1	ICT-Literacy WLE
ics3_sc2	ICT-Literacy SE of WLE

Exemplary data snapshot

ID_t	wave_w1	wave_w5	mas1_sc1	mas1_sc2
7006571	1	1	0.60330	0.53224
7017435	1	1	0.60668	0.56995
7004087	1	1	1.00120	0.61115
7008067	1	1	1.87321	1.01867
7018240	1	1	1.30902	0.81613

File xTargetCompetencies contains data from competence assessments conducted. Scored item variables as well as scale variables are available in a cross-sectional format. Note that in wave 1 competence tests were conducted in paper-and-pencil mode in groups at the participating higher education institutions. The wave 5 test included a mode experiment with an individual online test on the one hand and three different modes applied in a group setting (conventional paper-based assessment, paper-based assessment with digital pens, and computer-based assessment). Please also note that data from the web-based assessment are not available yet.

Due to overlaps in survey periods in wave 12, data was also collected from target persons who did not participate in the last three CATI waves (including CATI in wave 12) and who are usually treated as final drop-outs.

Stata 36: Working with xTargetCompetencies (find R example here)

```
** open datafile
use ${datapath}/SC5_xTargetCompetencies_D_${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 mas1_sc1 mas1_sc2

** save a temporary datafile
tempfile tmp
save `tmp'

** and merge this to CohortProfile
use ${datapath}/SC5_CohortProfile_D_${version}.dta, clear
merge 1:1 ID_t wave using `tmp', nogen
```

4.5.37 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

157 / 2,859

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Exemplary variables

ID_t	ID target
wave	Wave
t514001	Satisfaction with life
t514010	Satisfaction with study/training/school
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	t514010	tm00013	tm00015
7005988	.	6	9	I was employed	no
7017212	.	8	6	I was employed	yes
7018502	.	5	7	I was employed	no
7019220	.	6	7	I was employed	no
7036766	.	9	8	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 37: 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 Special Types of Variables

5.1.1 Service Variables

field of study The variables `tg2416*` were edited due to discrepancies between subspells. Subjects are filled for the first explicit mention only, missing information was labeled accordingly.

Currently the code *-29 "Value from last-mentioned sub-episode"* describes two cases: missing information can be found in the previous sub-spell or in the previous spell (the latter means a person started a new study-episode but claims that the subject is still the same as in the previously recorded episode).

The missing code *-28 "Value from recruitment pTargetCATI"* denotes that the missing information can be found in the recruitment data in file `pTargetCATI`.

The service variables `tg2416*_g1` and `tg2417*` contain information on the respective field of study, thus the variables `tg24162_g1`, `tg24165_g1`, `tg24168_g1`, `tg24170_g1`–`tg24170_g5`, `tg24173_g1`–`tg24173_g5`, and `tg24176_g1`–`tg24176_g5` provide complete subject information for all study episodes. Working with the service variables is recommended.

type of higher education institution The variable `tg01003_g1` (*type of higher education institution*, two levels) is originally a part of the first wave recruitment information contained in dataset `pTargetCATI`. The variable `ts15201` (*type of vocational training program*, seventeen levels) is part of the core education questionnaire and is recorded for each educational spell; it is part of `spVocTrain`. The service variable `tg01003_ha` (*type of higher education institution*) provides an aggregated version of `ts15201` in `spVocTrain` partly using information from `tg01003_g1` for first wave spells, as seen in table 9.

intended vocational qualification Because of a programming error, variable `ts15221` (*Intended vocational qualification*) misleadingly contains information on the *achieved* vocational qualification in waves 9 and 10 for some respondents. To correct this mistake, the variable `ts15221_g1` (*Intended vocational qualification, revised*) was introduced. This variable contains the correct information on the intended vocational qualification for all target persons in the data file and for all subspells of a vocational episode. The information on the intended vocational qualification for wave 1 was collected in the initial questionnaire and is stored in variable `tg02001` in `pTargetCATI`. This variable was used to provide the information in `ts15221_g1` in spVocTrain for wave 1.

Table 9: Harmonization of type of higher education institution

	tg01003_ha/tg01003_g1		ts15201
1	University of applied sciences (incl. Berufsakademie/cooperative state university)	7	Degree course at a Berufsakademie/cooperative state university
		8	Degree course at a college of public administration
		9	Degree course at a university of applied sciences (not a college of public administration)
2	University	10	Degree course at a university, including college of education, art college, music college

vocational education history In the telephone interview following wave 1 (i. e., in wave 3, 5, or 7, depending on panel participation), the vocational education history from winter term 2010/2011 onwards was newly collected with an improved survey instrument. This has led to duplicate and/or right-censored episodes in the dataset `spVocTrain`. In order to deal with those episodes, the variable `tx20100` was introduced to give a recommendation which episodes should be used for analyses. The rule applies that episodes from wave 1 are always recommended when the start date lies at or before the beginning of the first study episode of the winter term 2010/11. Episodes from wave 1 are never recommended when the start date lies after the beginning of the first study episode of the winter term 2010/11.

5.1.2 Auxiliary variables

Additionally to the reported information, auxiliary variables are generated automatically during the course of an interview. They are used to manage the interview process and to ensure that the questions are addressed adequately for each interviewee. Part of this supplementary stored information is useful when analysing the data. Therefore some of these automatically generated variables are released in different data sets. Auxiliary variables can be identified by their variable label, which begins with "Auxiliary variable:" and indicates, that it is not a recorded but an automatically generated information about the target person.

5.1.3 Version variables

It rarely happens that errors in the programming of the questionnaire appear during the field time, which could jeopardize the correct execution of the interview. In these cases, error corrections in the field are required. This information is stored in so-called version variables, so that it is documented later on in the data which target persons have received which application version of the instrument. One can identify such variables by their variable name starting

with `Version_`. A description of the error correction can be called up using the command `infoquery var` in Stata.

5.1.4 Preload Variables

In order to disburden the process of the online survey, information from prior waves is used to guide through the survey. It is important to note that only information from CATI waves is used to generate this preload information. Since some information in the online waves is only updated when a target person reports changes since the last CATI wave, some of these preloaded information is released in the data. You can identify the preloaded information by its label starting with "Preload:", indicating that this variable contains the information known at the time of the last CATI survey.

5.2 Coding field of study

5.2.1 Recruitment

data collection Information on field of study of the first study program in winter semester 2010/11 was collected mainly in PAPI and sometimes also in CATI mode (for information on sampling in SC5, see Aßmann et al., 2011, and Zinn et al., 2017). The data of the PAPI questionnaires were entered by the data collecting institute (infas) and delivered to NEPS. Information on the field of study was delivered to NEPS as original string variable.

coding Coding of field of study was done by the NEPS department *From Higher Education to the Labor Market* at DZHW Hannover (formerly HIS), based on data delivered by the data collecting institute (infas) from both modes (CATI and PAPI). The coding process faced a few challenges because the classification scheme used changed between recruitment and first wave data collection: sampling was based on the classification of 2009/10 while the coding of recruitment information was based on the classification of 2010/11 (see Statistisches Bundesamt, 2011).

Coding was done manually by occasionally using additional information when a decision could not be taken only based on the string variable.

classification used The classification used for coding the recruitment information on field of study is based on the Federal Statistical Office (Destatis) for the winter semester 2010/11 (Statistisches Bundesamt, 2011). Coding decisions can differ from Destatis recommendations for coding degree programs into fields of study due to individual decisions based on extensive research.

5.2.2 Panel Waves

data collection For higher education episodes reported after recruitment, the field of study has been recorded using lists – in CATI as well as in online surveys. In cases where interviewers were unable to fit a respondents answer into the respective list, the field of study has been recorded as an open string. Both in CATI and online panel waves, the lists are based on the destatis classification 2010/11 and the recruitment information.

To facilitate the allocation of respondents' answers, the CATI list has been continuously extended with supplementary information (based on open responses and changes in the academic landscape in Germany); the online list has remained the same.

Up until wave 13 subjective decisions in the maintenance of the CATI lists and technical restrictions have led to deviations from the original classification. In some cases, same subjects of study were assigned to *different* codes within the list. In other cases, multiple subjects were listed under the same code. The idea behind this was for the added subjects within the same code to serve as covariates, so interviewers could classify the respondents' answer into the *right* code in the list. Starting with wave 13, the CATI lists will only be extended in the sense that new subject names will be added to the existing subject groups corresponding to a code if those subject names are not already listed under another code. The allocation will follow the coding rules described below to ensure consistency and transparency. This way, the list for collecting the field of study will not be changed but will be enhanced over time. Starting with wave 14, online waves will use the CATI list of the previous CATI wave to harmonize the recording of fields of study in CATI and online mode.

coding Coding of open responses regarding field of study has been provided by the NEPS department *From Higher Education to the Labor Market* for all panel waves so far. Since SUF 6.0.0 all strings that have been coded once have been collected in a reference list with their corresponding code by the LIfBi Research Data Center to avoid inconsistencies. In the following waves, open strings have been matched with that list first and strings in the list automatically get assigned the same code. Open strings that have been reported for the first time were coded manually until SUF 9.0.0. Starting with SUF 10.0.0, coding has followed a set of standardized rules and the software CODI has been used.

classification used Data collection and coding of field of study largely follows the Destatis 2010/11 classification of fields of study.

derivation of SUF-variables In the Scientific Use File, several alternative variables containing information on the field of study are offered. Variables with the suffix `_g1R` and `_g2` contain aggregations of subjects according to the Destatis 2010/11 classification ("Studienbereich" and "Fächergruppe"), `_g3R`, `_g4R` and `_g5` contain derivations of the Destatis classification into different levels of the ISCED 97 classification (Statistisches Bundesamt, 2011). All derivations are based on a transcoding table provided by the Federal Statistical Office.

5.3 Coding of Higher Education Institutions

data collection In the initial questionnaire the information on Higher Education Institutions (HEI) was collected as an open string variable. In the following CATI and CAWI panel waves, new information on HEI was collected using lists. These lists are based on the destatis classification valid at the time of data collection (see further description below) and were extended to include coded open answers from further waves.

coding The open answers from the initial questionnaire were coded by the NEPS department *From Higher Education to the Labour market* using the Destatis classification of winter term 2010/2011. In CATI and CAWI panel waves, new open answers are coded using the extended list for data collection.

classification used Data collection and coding of HEI are based on the Destatis classifications of HEI since winter term 2010/2011. To match the current higher education area, the list used for data collection was updated annually according to the changes as presented in the Destatis classification. These changes include:

- Adding new HEI
- Deleting existing HEI
- Integration of one HEI into another
- Division in different locations/sites of HEI
- Renaming of HEI
- Renaming and changing of type of HEI
- Merging of locations
- Fusion of HEI

If the HEI codes stay the same, these changes have no consequences for data collection and coding. But there are modifications that result in new codes for existing institutions. Hence, it is possible that the institution ID (`ID_i`) in the data differs between respondents or between spells of one respondent, even though the respondents attend the same institution. Since winter term 2010/2011, these changes are listed in table table 10 (source: Destatis).

Table 10: Changes in HEI codes during survey

ID_i before	ID_i after	changed in
1003035	1002062	winter term 2013/14
1003089	1002366	winter term 2015/16
1003079	1002260	winter term 2016/17
1003142	1002987	winter term 2017/18
1003105	1002207	winter term 2018/19
1003137	1002165	winter term 2018/19

5.4 Special features of interruption episodes in spVocTrain

There are no sensible harmonization rules for interruption episodes. In the data, interruption episodes are filed in wide format (first interruption `_w1`, second interruption `_w2`, and third interruption `_w3`). The variables for a first / second / third interruption are being harmonized. However, these do not necessarily correspond between subspells. A second (persistent) break in the first wave may correspond to a first break in the second wave. For example:

subspell 1 two interruptions; the first interruption episode is completed; the second interruption episode is right-censored because it continues at the point of the survey

subspell 2 the continuing interruption is completed and stored in the variables for a first interruption

In addition, there was no range definition for the first interruption. The stimulus in the question was designed to report only interruptions since the last interview date, but if one is asked about the time of the first interruption in a course of study, then a wide range of information is possible. The target person could even think that the start time of a first of several interruption episodes is meant (beginning of the first interruption in the first subspell). This is why there is a small percentage in the data that reports a start date of the interruption before the interview date.

5.5 Teacher Education Students and Teachers

The sample of Starting Cohort 5 includes an oversample of teacher education students (see section 2.2 for further information). Since wave 8, the survey program is supplemented with specific questions for all (prospective) teachers – whether or not they belong to the oversample or basis sample. The participants who belong to the oversample can be identified by the variable `tx80121` (*Sample: oversample of teacher education students*) stored in `CohortProfile`. This information is important for analyzing selected waves. Because of funding issues the

oversample could not be invited to take part in wave 7 but remained in the sample as temporary dropouts. In wave 14 the oversample did not answer the questions regarding *job requirements* (`pTargetCAWI`: `tg781*; tg782*; tg783*; tg784*`) due to survey methodological reasons.

In the first panel wave, information whether a participant belongs to the group of teacher education students was collected in two different surveys – the initial paper and pencil questionnaire and the first telephone interview. Because of different question wording between the two surveys, this information is stored in different variables in `pTargetCATI`. Generally speaking there are two types of variables, described below, that help identify if a participants course of study leads to a teaching degree. First, there are those variables containing information about whether someone's degree is a teaching degree (intended teaching degree) and second, there are variables containing information about the type of teaching degree that is intended (type of teaching degree). Some of these variables contain only information about the study course of the first panel wave and others are applicable for all information about teaching degrees, irrespective of the panel wave it was collected. That helps to identify survey participants who entered the study with a non-teaching degree course of study and changed into a teaching degree course of study later on.

Identify intended teaching degree Variable `tg02001` (*Intended degree WT 2010 (PAPI questionnaire)*) contains information collected in the initial questionnaire. The variable distinguishes several degree types, including those that lead to a teaching degree in German teacher education (e.g., bachelor's degree, state examination). So-called polyvalent Bachelor courses with the option of specializing in teacher education are subsumed under the category "Bachelor (not in teaching)". In variable `tg02001_g1` (*Intended degree WT 2010 (incl. polyvalent Bachelor; PAPI questionnaire))*, this specific potential pathway to a teaching degree is recorded separately in the category "Polyvalent Bachelor with a teaching option".

In the data of the telephone interviews in `spVocTrain`, teacher education students can be identified by the variable `ts15221_g1` (*Intended vocational qualification, revised*) in combination with `tg24201` (*Intended teaching degree*). In contrast to the initial PAPI questionnaire, participants who report a higher education episode and want to obtain a non-teaching degree are asked whether they are studying with the aim of becoming a teacher. This new question was introduced because it is possible to be enrolled in a non-teaching degree course and to decide later for a teaching degree. To identify teacher education students in the first study episode of the winter term 2010/2011, additional information on the type of episode (variable `h_aktstu`) has to be used. For ease of use, a new service variable `tg02001_ha` (*Intended degree WT 2010 (start of study; CATI and PAPI questionnaire)*) was generated and introduced in pTargetCATI (see below).

`tg24201_g1` (*Intended teaching degree WT 2010 (start of study; CATI)*) contains information on whether the first study program in winter term 2010 was started with the aim of becoming a teacher. The information comes from the first telephone interview (variable `tg24201 if h_aktstu=1 and wave=1`). Because of different question word-

ing, the variable differs from tg02001 (*Intended degree WT 2010 (PAPI questionnaire)*). tg24201_g1 is part of pTargetCATI.

tg02001_ha (*Intended degree WT 2010 (start of study; CATI and PAPI questionnaire)*) combines information on the intended degree collected in the first PAPI questionnaire and the first telephone interview. It refers to the first study program in winter term 2010 and updates the variable tg02001 (*Intended degree WT 2010 (PAPI questionnaire)*) with information whether a teaching degree is intended, collected during the first telephone interview (tg24201). tg02001_ha is part of pTargetCATI.

Coding of type of (intended) teaching degree Data on the type of intended teaching degree were collected using an open-ended question. The coding of the answers is based on the classification of teaching careers proposed by the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Kultusministerkonferenz/KMK) but does not distinguish teacher education programs encompassing several levels. In case that more than one type of teaching degree or a program that span several levels was mentioned without specifying a main focus, the answer was coded under the highest level.

The corresponding variables are tg51420_g1 (*Type of intended teaching degree (differentiated; CAWI)*) in pTargetCAWI and tg24202_g1 (*Type of intended teaching degree (differentiated; CATI)*), tg24202_ha (*Type of intended teaching degree WT 2010 (start of study; CATI)*), and tg03001_g2 (*Type of intended teaching degree WT 2010 (differentiated; PAPI questionnaire)*) – all included in pTargetCATI. tg24202_g2 (*Type of intended teaching degree WT 2010 (start of study; CATI)*) contains information on the type of the intended teaching degree and refers to the first study program in winter term 2010. The information comes from the first telephone interview (variable tg24202_g1, if h_aktstu=1 and wave=1). tg24202_g2 is part of pTargetCATI.

Important notes on auxiliary variables The auxiliary variables tg60011 (wave 8), tg60014 (wave 9), tg60015 (wave 10), tg60016 (wave 13), have been generated to navigate through the survey. Please use these variables only to get a first overview of the data. Use the original episode files for analyses!

tg60012 and tg60013 do not only include information on the phase of teacher education a participant is pursuing or has completed but also on being employed as teacher. In addition, information on intentions are taken into account: Participants who have completed the second phase of teacher training (preparatory service) but are not yet employed as a teacher are asked whether they intend to work as a teacher. Respondents who have completed the first phase of teacher education at universities or equivalent institutions but have not yet started the second phase are asked whether they want to complete preparatory service. These two auxiliary variables are available from wave 11 onwards and are part of either pTargetCATI (tg60013) or pTargetCAWI (tg60012). Since variable tg60013 is used for guidance during the interview, there have been adjustments in this variable in wave 13 against wave 12 to ensure a stricter filtering into certain

teacher-related questions. It is now achieved that target persons who only have a qualification in a teacher-related bachelor's degree, but no longer intend to study a teacher-related master's degree, will no longer receive the teaching-related questions (as was still the case in wave 12). From wave 16 on there was a new answering category added so that variable `tg60013` contains now information about actual interrupted teaching employment episodes (*"6 = interrupted employment as a teacher (e.g. due to parental leave)"*). Survey participants with this status category are presented a reduced teacher context questionnaire. All these successive adjustments result in the existence of different versions of this variable (`tg60013_v1` in wave 12 vs. `tg60013_v2` in wave 13 and wave 15 vs. `tg60013` from wave 16 onwards) in the dataset `pTargetCAWI`. In addition, there is a correction to wave 15 data in `tg60013_g1v2`. For analyses in wave 13 and 15, please use this version variable instead of `tg60013_v2`. Since wave 14 the dataset `pTargetCAWI` contains an additional auxiliary variable to state the actual phase of teacher education of each participant (`tg60017`). The only difference to variable `tg60012` is that those participants who denied or revoked a teacher (education) context later in the questionnaire are now transferred to category *"0 = no teaching reference or status unknown"*.

5.6 Wave Specific Issues

The 2018 online survey (wave 14) experimented with various incentives. 50% of the sample was still offered to take part in a lottery. 25% of the sample was offered a cash incentive of EUR 10 and 25% of the sample had the choice between participating in the lottery or a cash incentive of EUR 10. The distribution across the three incentive groups was random. The assignment to the different incentive groups is documented in the variable `tg59000` (*Auxiliary Variable: Assignment B139 incentive group*) in `pTargetCAWI`. For further information see the respective field report at

→ `www.neps-data.de > Data Center > Data and Documentation`
`> Starting Cohort First-Year Students > Documentation`

For project-related reasons, the field period of the 2021 telephone survey for the study participants from the teaching oversample ended on June 19, 2021, and the teaching-specific survey program was no longer used after that date.

A References

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and Solutions (H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice, Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)*, *14*, 51–65. https://doi.org/10.1007/s11618-011-0181-8

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*.

Dahm, G. (2014). *Starting Cohort 5 - Dokumentation der Variable tg24150_g2 "NTS" (Nicht-traditionelle Studierende)* (DZHW: Data Manual). DZHW - Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH.

FDZ-LIfBi. (2023). *Data Manual NEPS Starting Cohort 5–First-Year Students, From Higher Education to the Labor Market, Scientific Use File Version 18.0.0*. 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.

Lauterbach, O. (2015). *Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des Nationalen Bildungspanels - Technischer Bericht* (NEPS Working Paper No. 51). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

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. (2023-a). *National Educational Panel Study, Scientific Use File of Starting Cohort First-Year Students*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC5:18.0.0.

NEPS Network. (2023-b). *Starting Cohort 5: First-Year Students (SC5), Wave 18, Questionnaires (SUF Version 18.0.0)*. 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.

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.

Statistisches Bundesamt (Ed.). (2011). *Bildung und Kultur. Fachserie 11 Reihe 4.1 - Studierende an Hochschulen. Wintersemester 2010/2011*.

Weber, A. (2014). *Data Manual: Starting Cohort 5 - Context Data*. DZHW. Hannover.

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.

Zinn, S., Steinhauer, H. W., & Aßmann, C. (2017). *Samples, Weights, and Nonresponse: the Student Sample of the National Educational Panel Study (Wave 1 to 8)* (NEPS Survey Paper No. 18). 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  <- "5"

# version of this Scientific Use File
version <- "18-0-0"
```

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 those libraries installed on your computer, you can easily do so by executing the command `install.packages("readstata13")` from inside R.

R 38: 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 39: Working with Basics

```
'** import the data files'
CohortProfile =
        read.dta13("SC5_CohortProfile_D_version.dta",
         convert.factors = T)

Basics =
        read.dta13("SC5_Basics_D_version.dta",
        convert.factors = T)

'** merge the data from Basics, enhancing every entry in CohortProfile'
CohortProfile = merge(CohortProfile, Basics, by = "ID_t", all = TRUE)
#The option all = TRUE makes sure that both, matched AND unmatched cases are kept
  during the merging process

'** tabulate gender by wave'
addmargins(table(Data$wave, Data$t700001))
```

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 CohortProfile

```
'** import the data file'
CohortProfile =
        read.dta13("SC5_CohortProfile_D_version.dta",
        convert.factors = T)

'** how many different respondents are there?'
length(unique(CohortProfile$ID_t))
#number of distinct ID_t

'** respondents in each wave'
cbind(addmargins(table(CohortProfile$wave)),
        addmargins(prop.table(table(CohortProfile$wave))))


'** check participation status by wave'
cbind(addmargins(table(CohortProfile$wave, CohortProfile$tx80220)))
```

R 42: Working with Education

```
'** we want to merge the school type from spSchool to this datafile.
 ** For this to work, we first have to prepare spSchool and keep only
 ** harmonized episodes (subspell == 0)'
spSchool =
        read.dta13("SC5_spSchool_D_version.dta",
        convert.factors = T)

spSchool = subset(spSchool, spSchool$subspell ==  0)

'** open the Education data file'
Education =
        read.dta13("SC5_Education_D_version.dta",
        convert.factors = T)

'** check which spell modules you can merge to this file'
table(Education$tx28100)

'** 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.
#If there are duplicates this command returns the index of the first duplicate


'** merge spSchool to Education'
#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:
Education = transform(merge(
  x = cbind(Education,source = "master"),
  #x contains the Education data set plus one extra column "source",
  #where source = "master"
  y = cbind(spSchool[,c("ID_t", "splink", "ts11204")],source = "using"),
  # y contains only the columns ID_t, splink and ts11204 from spSchool
  # 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
)


'** see that this only added information to the rows corresponding to spSchool'
cbind(addmargins(table(Education$tx28100, Education$source)))
```

R 43: Working with MethodsCATI

```
'** import the data file'
MethodsCATI =
        read.dta13("SC5_MethodsCATI_D_version.dta",
        convert.factors = T)

'** check out participation status by wave'
cbind(addmargins(table(MethodsCATI$wave, MethodsCATI$tx80220)))

'** how many different interviewers did CATI surveys?'
length(unique(MethodsCATI$ID_int))

'** create one single variable containing the interview date'
Sys.setlocale("LC_TIME", "C")
#turns off the location-specific language, so that the english months are recognized.

MethodsCATI$intdate  =
  as.Date(paste(MethodsCATI$intm, MethodsCATI$intd, MethodsCATI$inty, sep = '-'),
        "%B-%d-%Y")
#binds the three columns "intm", "intd" and "inty" into one new column "intdate"

head(MethodsCATI[c("intd", "intm", "inty", "intdate")], 10)
#displays first 10 rows of intd, intm, inty and intdate
```

R 44: Working with MethodsCompetencies

```
'** open the data file'
MethodsCompetencies =
        read.dta13("SC5_MethodsCompetencies_D_version.dta",
        convert.factors = T)


'** how many respondents have been tested together in a group'
MethodsCompetencies = within(MethodsCompetencies,{
  groupsize = ave(ID_tg, ID_tg, FUN = length)})
#creates a new variable "groupsize" and counts the observations in each ID_tg group

#Problem: NEPS-Missings are also counted as regular values and summirized in groups
for (i in 1:length(MethodsCompetencies$ID_tg)) {
  if(!is.na(MethodsCompetencies$ID_tg[i]) & MethodsCompetencies$ID_tg[i] < 0){
    MethodsCompetencies$groupsize[i] = NA
    #sets all observations to NA for which ID_tg < 0 (here -55 and -54)
  }
}

summary(MethodsCompetencies$groupsize)
#displays Min, Max and Mean for "groupsize"
sd(MethodsCompetencies$groupsize, na.rm = TRUE)
#displays Std.Dev. for "groupsize"
length(MethodsCompetencies$groupsize[!is.na(MethodsCompetencies$groupsize)])
#displays the number of observations in "groupsize" without NA
```

```
'** create duration of math test'
for (t in names(MethodsCompetencies[,c(38, 39)])) {
# run over columns 38 and 39 (variables tx80603 and tx80804)
  for (i in 1:length(MethodsCompetencies[[t]])) {
      #runs over every single observation
    if(nchar(MethodsCompetencies[[t]][i]) ==   3 & MethodsCompetencies[[t]][i] > 0) {
      #if the observation length is 3 and positive (e.g., "923", but not "-54")
      MethodsCompetencies[[t]][i] = paste("0", MethodsCompetencies[[t]][i], sep = "")
      #adds a leading 0 character, such that 923 becomes 0923
    }
  }
}

install.packages("chron")
library(chron)
#package for creating chronological objects

for (i in names(MethodsCompetencies[,c(38, 39)])){
  MethodsCompetencies[[paste(i, 't', sep = "_")]] =
    #creates new variables tx80603_t and tx80604_t
  times((strftime(strptime(MethodsCompetencies[[i]], format = "%H%M"),"%H:%M:%S")))
    #assigns the values from tx80603 and tx80604 in time format to them
}

MethodsCompetencies$duration =
        MethodsCompetencies$tx80604_t - MethodsCompetencies$tx80603_t
#creates a new variable "duration", subtracting start time from end time

summary(MethodsCompetencies$duration)
#displays Min, Max and Mean for "duration" in time format
mean(MethodsCompetencies$duration) * 60 * 24
#displays the mean in minutes format
#one unit equals one day, thus it has to be multiplied by 60 minutes and 24 hours

sd(MethodsCompetencies$duration, na.rm = TRUE) * 60 * 24
#displays Std.Dev. for "duration" in minutes format
times(sd(MethodsCompetencies$duration, na.rm = TRUE))
#displays Std.Dev. in time format

length(MethodsCompetencies$duration[!is.na(MethodsCompetencies$duration)])
#displays the number of observations in "duration" without NA
```

R 45: Working with pTargetCATI

```
'** open the CohortProfile dataset'
CohortProfile =
        read.dta13("SC5_CohortProfile_D_version.dta",
        convert.factors = T)

'** merge some variable from pTargetCATI'

pTargetCATI =
```

```
        read.dta13("SC5_pTargetCATI_D_version.dta",
        convert.factors = T)
#imports the pTargetCATI dataset

CohortProfile =
        merge(x = CohortProfile,
        y = pTargetCATI[,c("ID_t",  "wave", "t400500_g1", "t525204")],
        by = c("ID_t", "wave"), all.x = TRUE)
#merges only variables "t400500_g1" and "t525204" from pTargetCATI to CohortProfile

'** note: this information is available only in waves which have surveyed the topic'
addmargins(table(CohortProfile$wave, CohortProfile$t400500_g1))

'** if it makes sense, you can copy this information to cells of other waves.
 ** This copies information downwards (i.e., to late waves), unless a new
 ** value has been reported (which is usually what you want in a panel study'
for (i in 2:length(CohortProfile$ID_t)) {
  if(CohortProfile$ID_t[i] ==  CohortProfile$ID_t[i-1]) {
    if(is.na(CohortProfile$t400500_g1[i]) |
      CohortProfile$t400500_g1[i] ==  "Missing by design") {
      CohortProfile$t400500_g1[i] = CohortProfile$t400500_g1[i-1]
    }
  }
}

addmargins(table(CohortProfile$wave, CohortProfile$t400500_g1))
```

R 46: Working with pTargetCAWI

```
'** open the pTargetCAWI dataset'
pTargetCAWI = read.dta13("SC5_pTargetCAWI_D_version.dta", convert.factors = T)

'** only keep single variables and IDs'
pTargetCAWI = subset(pTargetCAWI, select = c(ID_t, wave, t289902))

'** suppose you want to know if somebody ever lived with roommates.
 ** t289902 == "Specified" if there has been a roommate,
 ** and t289902 == "Not specified" otherwise. The maximum of
 ** this expression over waves results in 1 if any wave ever evaluated to true,
 ** and 0 otherwise.'
for (i in 1:length(pTargetCAWI$ID_t)){
  if(pTargetCAWI$t289902[i] == "Specified")pTargetCAWI$roommate[i] = 1
        else pTargetCAWI$roommate[i] = 0
}

pTargetCAWI = within(pTargetCAWI, {roommate = ave(roommate, ID_t, FUN = max)})
#for every ID_t with at least one roommate == 1, all other roommate observations
#are also replaced by 1 within this ID_t.

'** only keep this variable; as all waves contain the same information, we
 ** can fall back to cross-sectional structure'
pTargetCAWI = subset(pTargetCAWI, select = c(ID_t, roommate))
pTargetCAWI = pTargetCAWI[!duplicated(pTargetCAWI),]
```

```
'** finally, open CohortProfile and merge this variable'
CohortProfile = read.dta13("SC5_CohortProfile_D_version.dta", convert.factors = T)
CohortProfile = merge(CohortProfile, pTargetCAWI, by = c("ID_t"), all = TRUE)
addmargins(table(CohortProfile$wave, CohortProfile$roommate))
```

R 47: 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 48: Working with spChild

```
'** open the data file'
spChild = read.dta13("SC5_spChild_D_version.dta", convert.factors = T)

'** 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'
levels(spChild$ts3320m)[levels(spChild$ts3320m) == "Beginning of the year/winter"] =
  "January"
levels(spChild$ts3320m)[levels(spChild$ts3320m) == "Spring/Easter"] = "April"
levels(spChild$ts3320m)[levels(spChild$ts3320m) == "Mid-Year/Summer"] = "July"
```

```
levels(spChild$ts3320m)[levels(spChild$ts3320m) == "Fall"] = "October"
levels(spChild$ts3320m)[levels(spChild$ts3320m) == "End of year"] = "December"

'** compute the age of 'ones children today
 ** first, create a date of the birth variables'
spChild$ts3320m = match(spChild$ts3320m, month.name)

install.packages("zoo")
library(zoo)
#the zoo package is needed to transform time data

#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 49: Working with spChildCohab

```
'** open the data file'
spChildCohab = read.dta13("SC5_spChildCohab_D_version.dta", convert.factors = T)


'** only keep full or harmonized episodes'
spChildCohab = subset(spChildCohab, spChildCohab$subspell ==  0)

'** recode rough values (e.g., end of year) to real months'
for (i in names(spChildCohab[c(16, 18)])){
        #run over the variables ts3331m and ts3332m in columns 16 and  18
  levels(spChildCohab[[i]])[levels(spChildCohab[[i]]) == "Beginning of the year/
    winter"] = "January"
  levels(spChildCohab[[i]])[levels(spChildCohab[[i]]) == "Spring/Easter"] = "April"
  levels(spChildCohab[[i]])[levels(spChildCohab[[i]]) == "Mid-Year/Summer"] = "July"
  levels(spChildCohab[[i]])[levels(spChildCohab[[i]]) == "Fall"] = "October"
  levels(spChildCohab[[i]])[levels(spChildCohab[[i]]) == "End of year"] = "December"
}

'** generate the following durations in months:
 * a) the total duration of a cohabitation episode'
for (i in names(spChildCohab[c(16, 18)])) {
  spChildCohab[[i]] = match(spChildCohab[[i]], month.name)
  #transforms month names into month numbers
}
```

```
install.packages("zoo")
library(zoo)
#the zoo package is needed to transform time data

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
 ** which gives you a cross-sectional display of cohabitation time per respondent'
spChildCohab = subset(spChildCohab, select = c("ID_t", "total_duration_per_target"))
spChildCohab = spChildCohab[!duplicated(spChildCohab),]
```

R 50: Working with spCourses

```
'** open the data file'
spCourses = read.dta13("SC5_spCourses_D_version.dta", convert.factors = T)

'** check which modules provided course information'
cbind(addmargins(table(spCourses$sptype)))

'** only keep courses from employment spells'
spCourses = subset(spCourses, spCourses$sptype ==  "Emp")

'** open the employment module'
spEmp = read.dta13("SC5_spEmp_D_version.dta", convert.factors = T)

'** merge spCourses to spEmp
 ** note that this is an m:1 merge, as there are still subspells in spEmp'
#Since the variable tx80211 is in both data sets spCourses AND spEmp
intersect(names(spCourses), names(spEmp))
#and since the variable is not one of the merging variables, both versions
#are contained in the new merged data set as tx80211.x and tx80211.y.

#To avoid that there are two possibilities:
```

```
#1. You can include the variable in the merging process by:
spEmp =
  merge(spEmp, spCourses, by = c("ID_t", "wave", "splink", "tx80211"), all.x = TRUE)
# In that case the version from the master data set, here spEmp, is kept

#OR

#2. If you'd like to compare the both versions first, you can merge the
#data sets as usual by:
spEmp =
  merge(spEmp, spCourses, by = c("ID_t", "wave", "splink"), all.x = TRUE)

#compare the two versions of the variable tx80211 by:
addmargins(table(spEmp$tx80211.x, spEmp$tx80211.y))

#and then drop one of the variables by:
spEmp$tx80211.y = NULL

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

R 51: Working with spEmp

```
'** open the data file'
spEmp = read.dta13("SC5_spEmp_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spEmp = subset(spEmp, spEmp$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 52: Working with spFurtherEdu1

```
'** open the datafile'
spFurtherEdu1 = read.dta13("SC5_spFurtherEdu1_D_version.dta", convert.factors = T)

'** 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[,3:11]),
                #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.dta13("SC5_CohortProfile_D_version.dta", convert.factors = T)


'** 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 53: Working with spFurtherEdu2

```
'** Two possibilities to use spFurtherEdu2'

'-------------------------------------------------------'
'** A) Merge data to spCourses'

'** open spCourses datafile'
spCourses = read.dta13("SC5_spCourses_D_version.dta", convert.factors = T)

'** one row contains information for up to three courses.
** To make merging possible, you first have to reshape the datafile
** so one row contains only one course'
spCourses = reshape(data = spCourses,
                    # data in wide format
                    idvar = c("ID_t","wave","splink"),
                    #idvar is/are the variable/s that need/s to be left unaltered
                    varying = c("course_w1","course_w2","course_w3"),
                    #varying are the variables that need to be converted from
                    #wide to long
                    v.names = c("course"),
                    #v.names defines the name of the variable in that the in
                    #varying defined variables are summarized
                    times = c(1,2,3),
                    #new variable "time" is created with levels 1, 2 and 3
                    #for the three courses
                    new.row.names = 1:100000,
                    #sets row names as numeric
                    direction = "long"
                    ##direction is to which format the data needs to be transformed
                    )

names(spCourses)[names(spCourses) == "time"] <- "course_nr"
#renames the variable "time" to "course_nr"


'** merge spFurtherEdu2 using ID_t and course'
#open spFurtherEdu2 datafile
spFurtherEdu2 = read.dta13("SC5_spFurtherEdu2_D_version.dta", convert.factors = T)

intersect(names(spCourses), names(spFurtherEdu2))
#common variables in the both data sets are "ID_t", "wave", "tx80211" and "course"
#Since the variables "wave" and "tx80211" are not part of the merging process,
#both versions are contained in the new merged data set
#as wave.x/wave.y and tx80211.x/tx80211.y.

'**To avoid that, there are two merging options:'
#1. You can include the variables in the merging process by:
spCourses =
        merge(spCourses, spFurtherEdu2,
        by = c("ID_t", "course", "wave", "tx80211"), all.x = TRUE)
```

```
#In that case the versions from the master data set are kept (wave.x and tx80211.x)

#OR

#2. If you'd like to compare the both versions first,
#you can merge the data sets as usual by:
spCourses = merge(spCourses, spFurtherEdu2, by = c("ID_t", "course"), all.x = TRUE)

#compare the two versions of the variables by:
addmargins(table(spCourses$wave.x, spCourses$wave.y))
addmargins(table(spCourses$tx80211.x, spCourses$tx80211.y))

#and then drop one of the versions by:
spCourses$wave.y = NULL
spCourses$tx80211.y = NULL
'-------------------------------------------------------'


'-------------------------------------------------------'
'** B) merge to spFurtherEdu1'

'** open spFurtherEdu1 and FurtherEdu2 datafiles'
spFurtherEdu1 = read.dta13("SC5_spFurtherEdu1_D_version.dta", convert.factors = T)
spFurtherEdu2 = read.dta13("SC5_spFurtherEdu2_D_version.dta", convert.factors = T)

'** merge spFurtherEdu2 using ID_t and courses'

intersect(names(spFurtherEdu1), names(spFurtherEdu2))
#common variables in the both data sets are "ID_t", "wave", "course" and "tx80211"
#Since the variables "wave" and "tx80211" are not part of the merging process,
#both versions are contained in the new merged data set
#as wave.x/wave.y and tx80211.x/tx80211.y.

'**To avoid that, there are two merging options:'
#1. You can include the variables in the merging process by:
spFurtherEdu1 =
        merge(spFurtherEdu1, spFurtherEdu2,
        by = c("ID_t", "course", "wave", "tx80211"), all.x = TRUE)
#In that case the versions from the master data set are kept (wave.x and tx80211.x)

#OR

#2. If you'd like to compare the both versions first,
#you can merge the data sets as usual by:
spFurtherEdu1 =
        merge(spFurtherEdu1, spFurtherEdu2,
        by = c("ID_t", "course"), all.x = TRUE)

#compare the two versions of the variables by:
addmargins(table(spFurtherEdu1$wave.x, spFurtherEdu1$wave.y))
addmargins(table(spFurtherEdu1$tx80211.x, spFurtherEdu1$tx80211.y))

#and then drop one of the versions by:
spFurtherEdu1$wave.y = NULL
```

```
spFurtherEdu1$tx80211.y = NULL
'-------------------------------------------------------'
```

R 54: Working with spGap

```
'** open the data file'
spGap = read.dta13("SC5_spGap_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spGap = subset(spGap, spGap$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 55: Working with spInternship

```
'** open the data file'
spInternship = read.dta13("SC5_spInternship_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spInternship = subset(spInternship, spInternship$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** merge spInternship 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(spInternship,source = "using"),
  #y contains the spInternship 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 spInternship
#check before merging by: intersect(names(Biography), names(spInternship))
#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 56: Working with spMilitary

```
'** open the data file'
```

```
spMilitary = read.dta13("SC5_spMilitary_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spMilitary = subset(spMilitary, spMilitary$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 57: Working with spParLeave

```
'** open the data file'
spParLeave = read.dta13("SC5_spParLeave_D_version.dta", convert.factors = T)
```

```
'** only keep full or harmonized episodes'
spParLeave = subset(spParLeave, spParLeave$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 58: Working with spPartner

```
'** open the data file'
spPartner = read.dta13("SC5_spPartner_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spPartner = subset(spPartner, spPartner$subspell ==  0)
```

```
'** to find out if a respondent has ever been lived together with a partner,
 ** you could'
cbind(addmargins(table(spPartner$t733030)),
        addmargins(prop.table(table(spPartner$t733030))))
```

R 59: Working with spSchool

```
'** open the data file'
spSchool = read.dta13("SC5_spSchool_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spSchool = subset(spSchool, spSchool$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 60: 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'
#open pTargetCATI
pTargetCATI = read.dta13("SC5_pTargetCATI_D_version.dta", convert.factors = T)

#display value labels
levels(pTargetCATI$wave)

#keep only the first wave as this data is time-invariant
pTargetCATI =
        subset(pTargetCATI, pTargetCATI$wave ==  "2010/2011 (CATI+competencies)")

#keep only ID_t, t70000m and t70000y from pTarget
pTargetCATI =
        subset(pTargetCATI, select = c("ID_t", "t70000m", "t70000y"))


'** now, open the data file spSchoolExtExam'
spSchoolExtExam =
  read.dta13("SC5_spSchoolExtExam_D_version.dta", convert.factors = T)


'** merge the previously extracted birth dates in pTargetCATI to spSchoolExtExam'
spSchoolExtExam = merge(spSchoolExtExam, pTargetCATI, by = c("ID_t"), all.x = TRUE)


'** recode the two date variables (year, month) into one:'

Sys.setlocale("LC_TIME", "C")
#turns off the location-specific language, such that the english month names
#are recognized as months.

spSchoolExtExam$ts1130m = match(spSchoolExtExam$ts1130m, month.name)
spSchoolExtExam$t70000m = match(spSchoolExtExam$t70000m, month.name)
#transforms month names into month numbers


install.packages("zoo")
library(zoo)
#the zoo package is needed to transform time data

spSchoolExtExam$exam_date =
        as.yearmon(paste(spSchoolExtExam$ts1130y, spSchoolExtExam$ts1130m), "%Y %m")
spSchoolExtExam$birth_date =
        as.yearmon(paste(spSchoolExtExam$t70000y, spSchoolExtExam$t70000m), "%Y %m")
```

```
#recode the two date variables (year, month) into one

'** 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),Freq = length(x)))
#displays mean and sd of age by school-leaving qualification

summary(spSchoolExtExam$age)
#display mean of age in general

sd(spSchoolExtExam$age, na.rm = TRUE)
#display sd of age in general
```

R 61: 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'
#open pTargetCATI
pTargetCATI = read.dta13("SC5_pTargetCATI_D_version.dta", convert.factors = T)

#display value labels
levels(pTargetCATI$wave)

#keep only the first wave as this data is time-invariant
pTargetCATI =
        subset(pTargetCATI, pTargetCATI$wave ==  "2010/2011 (CATI+competencies)")

#keep only ID_t, t70000m and t70000y from pTarget
pTargetCATI = subset(pTargetCATI, select = c("ID_t", "t70000m", "t70000y"))

'** now, open the data file spSibling'
spSibling = read.dta13("SC5_spSibling_D_version.dta", convert.factors = T)

'** merge the previously extracted birth dates in pTargetCATI to spSibling'
spSibling = merge(spSibling, pTargetCATI, by = c("ID_t"), all.x = TRUE)

'** recode the two date variables (year, month) into one:'
Sys.setlocale("LC_TIME", "C")
#turns off the location-specific language, such that the english month names are
  recognized as months.

spSibling$tg3270m = match(spSibling$tg3270m, month.name)
spSibling$t70000m = match(spSibling$t70000m, month.name)
#transforms month names into month numbers
```

```
install.packages("zoo")
library(zoo)
#the zoo package is needed to transform time data

spSibling$sibling_bdate =
        as.yearmon(paste(spSibling$tg3270y, spSibling$tg3270m), "%Y %m")
spSibling$target_bdate =
        as.yearmon(paste(spSibling$t70000y, spSibling$t70000m), "%Y %m")
#recode the two date variables (year, month) into one

'** 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 62: Working with spUnemp

```
'** open the data file'
```

```
spUnemp = read.dta13("SC5_spUnemp_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spUnemp = subset(spUnemp, spUnemp$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 63: 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'
#open pTargetCATI
pTargetCATI = read.dta13("SC5_pTargetCATI_D_version.dta", convert.factors = T)

#display value labels
levels(pTargetCATI$wave)

#keep only the first wave as this data is time-invariant
pTargetCATI =
        subset(pTargetCATI, pTargetCATI$wave ==  "2010/2011 (CATI+competencies)")

#keep only ID_t, t70000m and t70000y from pTarget
pTargetCATI = subset(pTargetCATI, select = c("ID_t", "t70000m", "t70000y"))

'** open the data file spVocExtExam'
spVocExtExam = read.dta13("SC5_spVocExtExam_D_version.dta", convert.factors = T)

'** merge the previously extracted birth dates in pTargetCATI to spVocExtExam'
spVocExtExam = merge(spVocExtExam, pTargetCATI, by = c("ID_t"), all.x = TRUE)

'** recode the two date variables (year, month) into one:'
Sys.setlocale("LC_TIME", "C")
#turns off the location-specific language, such that the english month names are
  recognized as months.

spVocExtExam$ts1530m = match(spVocExtExam$ts1530m, month.name)
spVocExtExam$t70000m = match(spVocExtExam$t70000m, month.name)
#transforms month names into month numbers

install.packages("zoo")
library(zoo)
#the zoo package is needed to transform time data

spVocExtExam$exam_date =
        as.yearmon(paste(spVocExtExam$ts1530y, spVocExtExam$ts1530m), "%Y %m")
spVocExtExam$birth_date =
        as.yearmon(paste(spVocExtExam$t70000y, spVocExtExam$t70000m), "%Y %m")
#recode the two date variables (year, month) into one

'** 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),Freq = length(x)))
#displays mean and sd of age by school-leaving qualification

summary(spVocExtExam$age)
#displays mean of age in general

sd(spVocExtExam$age, na.rm = TRUE)
#displays sd of age in general
```

R 64: Working with spVocPrep

```
'** open the data file'
spVocPrep = read.dta13("SC5_spVocPrep_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spVocPrep = subset(spVocPrep, spVocPrep$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 65: Working with spVocTrain

```
'** open the data file'
spVocTrain = read.dta13("SC5_spVocTrain_D_version.dta", convert.factors = T)

'** only keep full or harmonized episodes'
spVocTrain = subset(spVocTrain, spVocTrain$subspell ==  0)

'** open the Biography data file'
Biography = read.dta13("SC5_Biography_D_version.dta", convert.factors = T)

'** 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 66: Working with Weights

```
'** open the data file'
Weights = read.dta13("SC5_Weights_D_version.dta", convert.factors = T)

'** note that this file is cross-sectional,
 **although the weights seem to contain panel logic'
attr(Weights, "var.labels")

'** only keep weights corresponding to all waves'
Weights = subset(Weights, select = c(ID_t, w_t123456789))

'** create a "panel" logic, i.e., clone each row'
Weights = Weights[rep(seq_len(nrow(Weights)), each = 9),]

'** then create a wave variable'
Weights$wave = ave(Weights$ID_t, Weights$ID_t, FUN = seq_along)

'** open CohortProfile'
CohortProfile = read.dta13("SC5_CohortProfile_D_version.dta", convert.factors = T)

#value labels of "wave" in "CohortProfile" and "Weights"
#have to be synchronized before merging
levels((CohortProfile$wave))
levels(Weights$wave)

Weights$wave = as.factor(Weights$wave)
#sets "wave" in "Weights" as factor

for (i in 1:9) {
  levels(Weights$wave)[i] = levels(CohortProfile$wave)[i]
  #assigns the same value labels to "wave" in "Weights" as in "CohortProfile"
}

'** and merges Weights to CohortProfile'
CohortProfile = merge(CohortProfile, Weights, by = c("ID_t", "wave"), all = TRUE)

'** note that this weight is only nonzero if respondents participated in all waves'
with(subset(CohortProfile, w_t123456789 ! =  0), addmargins(table(wave, tx80220)))
```

R 67: Working with xInstitution

```
'** open datafile pTargetCATI'
pTargetCATI = read.dta13("SC5_pTargetCATI_D_version.dta", convert.factors = T)

'** copy the information from the first wave downwards for each target,
 ** unless a new value has been reported'
for (t in names(pTargetCATI[c("ID_i", "tg04001_g7")])) {
#run over variables ID_i and tg04001_g7
 for (i in 2:length(pTargetCATI$ID_t)) {
 #run over all observations
   if(pTargetCATI$ID_t[i] ==  pTargetCATI$ID_t[i-1]){
          #for the same ID_t, check...
     if(is.na(pTargetCATI[[t]][i]) | pTargetCATI[[t]][i] ==  "Missing by design"){
        #...whether missing value or -54(Missing by design)
```

```
        pTargetCATI[[t]][i] = pTargetCATI[[t]][i-1]
        #copy information downwards, unless a new value has been reported
      }
    }
  }
}

'** drop all observations where no satisfaction with studies was reported'
levels(pTargetCATI$t514008)

#remove observations with NA in t514008
pTargetCATI = pTargetCATI[!(is.na(pTargetCATI$t514008)),]

#remove observations with other missings in t514008
pTargetCATI = subset(pTargetCATI, !(t514008 ==  "Don't know"
                                  | t514008 ==  "Refused"
                                  | t514008 ==  "Does not apply"
                                  | t514008 ==  "Missing by design"))


'** some respondents reported satisfaction with studies in 7th and in 9th waves
 ** to keep the latest information, create a seq and a max variables'
pTargetCATI = within(pTargetCATI,{seq = ave(ID_t, ID_t, FUN = seq_along)})
pTargetCATI = within(pTargetCATI,{max = ave(ID_t, ID_t, FUN = length)})

'** only keep the latest reported iformation'
pTargetCATI =
        subset(pTargetCATI, pTargetCATI$seq ==  pTargetCATI$max)

'** only keep the variables relevant for the merge and the analysis'
pTargetCATI =
        subset(pTargetCATI, select = c("ID_t", "ID_i", "tg04001_g7", "t514008"))

'** merge two variables from xInstitution'

#open datafile xInstitution
xInstitution = read.dta13("SC5_xInstitution_O_version.dta", convert.factors = T)

#merge xInstitution to pTargetCATI
pTargetCATI =
  merge(x = pTargetCATI,
              y = xInstitution[,c("ID_i", "g04001_g7", "tg92601_R", "tg92104_O")],
              by = c("ID_i", "g04001_g7"), all.x = TRUE)

'** assuming that the less students at university the more intensive the support by
 ** the university staff per student and the more satisfied are students with their
 ** studies tabulate Satisfaction with studies by Students 2010 total
 ** note that the following analysis is feasible in both, RemoteNEPS and Onsite'
cbind(addmargins(table(pTargetCATI$t514008, pTargetCATI$tg92601_R)))
cbind(addmargins(prop.table(table(pTargetCATI$t514008, pTargetCATI$tg92601_R))))


'** assuming that students at excellence universities are more satisfied with
 ** their studies, tabulate the distribution of satisfaction by tg92104_O
```

```
 ** note that the following analysis is only feasible in the Onsite version of SUF,
 ** since the variable tg92104_O is anonymized in RemoteNEPS'
cbind(addmargins(table(pTargetCATI$t514008, pTargetCATI$tg92104_O)))
cbind(addmargins(prop.table(table(pTargetCATI$t514008, pTargetCATI$tg92104_O))))
```

R 68: 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 69: Working with xTargetCompetencies

```
'** open the data file xTargetCompetencies'
xTargetCompetencies =
        read.dta13("SC5_xTargetCompetencies_D_version.dta", convert.factors = T)

'** 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 competence testing has been conducted in multiple waves
 ** an indicator marks if a row contains information for a specific wave'
table(xTargetCompetencies$wave_w1)
table(xTargetCompetencies$wave_w5)
table(xTargetCompetencies$wave_w7)

'** 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.
 ** here, we focus on math competencies, that have been tested in wave 1.'
xTargetCompetencies$wave =
        rep(levels(CohortProfile$wave)[1],length(xTargetCompetencies$ID_t))
xTargetCompetencies$wave = as.factor(xTargetCompetencies$wave)

'** now, keep cases which did took part in the testing'
xTargetCompetencies = subset(xTargeCompetencies, wave_w1 ==  "ja")
```

```
'** and reduce the dataset to the relevant variables'
xTargetCompetencies =
        subset(xTargetCompetencies, select = c(ID_t, wave, mas1_sc1, mas1_sc2))

'** and merge this to CohortProfile'

#open the data file Cohort Profile
CohortProfile = read.dta13("SC5_CohortProfile_D_version.dta", convert.factors = T)

#look for common variables in both data sets
intersect(names(CohortProfile), names(xTargetCompetencies))

#merge CohortProfile with xTargetCompetencies
CohortProfile =
  merge(CohortProfile, xTargetCompetencies, by =  c("ID_t", "wave"), all = TRUE)
```

R 70: 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 5 - RELEASE NOTES a.k.a CHANGE LOG
** changes and updates for release NEPS SC5 18.0.0
** (doi:10.5157/NEPS:SC5:18.0.0)
**

==================================================
* Changes introduced to NEPS:SC5 by version 18.0.0 *
==================================================

General:
        - the information on the linkage with administrative data from the IAB in
            variables tx80130, tx80131, tx80132
                (TargetMethods) and tx80533 (CohortProfile) has been updated and now
                    refers to version
                "NEPS-SC5-ADIAB 7520 v1"

        - all missing values "-29 value from the last sub-episode" in all spellfiles
            have been replaced by the respective value;
                The data now also contains information that was not asked directly from
                     the respondent but was necessary for
                interview and filter control. These therefore represent the last known
                    value and can be used to track the filtering.

Deletion of variables:
        - some variables were completely filtered for cohort 5 and therefore these
            variables held no valid information. Due to that fact these items were
            deleted for this release:
                * pTargetCATI: tf13301 tf13302 tf13303 tf13304 tf13305 tf13306_O
                    tf13307 ts13288
                * spVocPrep: ts13104 ts13105_g1 ts13105_g2R ts13105_g3R ts13105_g4R
                    ts13202 ts13203 ts13204 ts13205 ts13206 ts13207 ts13208 ts13209
                    ts13210 ts13211 ts13212 ts13213 ts13214 ts13215 ts13281 ts13282
                    ts13285 ts13286 ts13287 ts13289 ts13290 ts13291 ts13292 ts13293
                    ts13294 ts13295 ts13296 ts13297 ts13298 ts13299
                * xTargetCORONA: tm00033 tm00034 tm00035 tm00036 tm00037 tm00038
                    tm00039 tm00040_O tm00054 tm00055 tm00060 tm00061 tm00062 tm00063
                    tm00064 tm00065 tm00066 tm00067 tm00068 tm00069 tm00070 tm00071_O
                    tm00072 tm00073 tm00074 tm00075 tm00076


==================================================
* Changes introduced to NEPS:SC5 by version 17.0.0 *
==================================================
xTargetCORONA:
        - pTargetCORONA was renamed to xTargetCORONA as no additional waves will be
            added to this dataset

pTargetCAWI:
        - the variable name tg69225 ("T - structure (class management): monitoring 5")
            was corrected to tg69725

xInstitution/spVocTrain/StudyStates:
```

```
        - type of university within ID_i was harmonized between datasets. In former
            releases information given by respondent concerning type of university was
             trusted without further checks

MethodsCATI:
        - data on interviewer characterics added for wave 16


known issues:

StudyStatesAddon:
        - this dataset will be possibly provided in another major release to analyze
            target persons with multiple episodes within waves

Basics:
        - use this dataset just to get an idea of the sample, please! Do not analyze
            data using this dataset!


=====================================================
* Changes introduced to NEPS:SC5 by version 16.0.0 *
=====================================================

in general:
- disagint has been extended for revoked episode in check module
- removed generated date variables (*12?_g1, *11?_g1) in spell datasets
- new variable names for intm inty intd corresponding to those in SC3 and SC4
- information taken from CATI/spVocTrain and CAWI can differ and show inconsistencies,
    this is due to the nature of survey data


New dataset added: spVocBreaks

        - data on vocational breaks were extracted from spVocTrain
        - break episodes were reshaped to long format
        - break episodes were cleaned (closing small gaps, drop breaks within breaks)


StudyStates:
        - added new variables (including CAWI-information)
                >> type of higher education institution (CATI/spVocTrain): tx92401
                    tx92402
                >> type of higher education institution (CAWI): tx92403
                >> Status of the course of study (CAWI): tx24101
                >> Higher education institution ID (CAWI): tx24013
                >> Highest vocational qualification (CAWI): tx15316


=====================================================
* Changes introduced to NEPS:SC5 by version 15.0.0 *
=====================================================


in general:
- supplemental meta information on several auxiliary variables was added
- A duplicate was removed from the data

New dataset added: StudyStates

StudyStates:
        - (experimental) dataset on states, breaks and changes of tertiary education
```

```
        - structure of dataset should stay that way, certain states of persons are not
            fully plausibilized yet!
        - dataset contains variable tx24000 to recommend usage of data for persons with
             less than two simultaneous episodes in any wave, this excludes 641 target
             persons.
          Multiple episodes in waves cause problems to generate states, change of
             subjects, study breaks etc.

pTargetCATI:
        -  versionized variable t514008_v1 for wave 7 had to be added. In this wave all
             target persons were asked about their satisfaction with the course of
            studies - no matter if they were actually studying at this time. Because
            of this, interviewers were instructed to tell non-students to choose the "
            does not apply"-button. From wave 9 on, non-students were filtered and
            interviewer instructions were removed.
        - job related information on social capital was implemented in wave 15 by
            mistake. After this problem has been recognized, the items were removed
            from the instrument. As only a part of the respondents answered these
            questions, this variables were removed for wave 15. These variables are:
            t324110 t32411k t32411l t32411m t32411o t32411p t32411q t32411n t32411r
            t32411s t32411b t32411d t32411e t32411c t325110 t32511k t32511l t32511m
            t32511o t32511p t32511q t32511n t32511r t32511s t32511b t32511d t32511e
            t32511c
        - variable t515039_g1 "Job characteristics: episode number of employment
            episode" was generated for wave 15 for merging information from spEmp to
            information on job characteristics in pTargetCATI

pTargetCORONA:
        - information on satisfaction with course of study, school or apprenticeship (
            variable t514010) is also available for those respondents who previously
            indicated that they were employed, although the response option "does not
            apply" was available; in these cases, it is unclear what respondents were
            referring to with their answer to the satisfaction question, so this
            variable should be treated with caution, depending on the research
            question
        - information on satisfaction with work (variable t514009) is also available
            for those respondents who previously indicated that they studied, were in
            vocational educational training or did nothing, although the response
            option "does not apply" was available; in these cases, it is unclear what
            respondents were referring to with their answer to the satisfaction
            question, so this variable should be treated with caution, depending on
            the research question
        - information on health-related limitations are also available in pTargetCORONA
             (variable t521055). However, some respondents indicated very strong or
            strong limitations; they stated that they are in good or very good
            physical and mental health, though. In these cases, it is unclear, how
            respondents interpreted the question regarding limitations in daily
            activities.

spPartner:
        - variables ts31226_g1 - ts31226_g16 "Partner: Profession" were edited. If the
            partner's profession did not change since the last interview (th32369 ==
            1), the code -29 "Value from the last sub-episode" was implemented.


====================================================
* Changes introduced to NEPS:SC5 by version 14.1.0 *
====================================================

New dataset added: pTargetCORONA

pTargetCATI:
```

```
        - added values to t751004_g* - variables for wave 12 and wave 13

=====================================================
* Changes introduced to NEPS:SC5 by version 14.0.0 *
=====================================================

General remarks:
        - some variable labels were corrected
        - supplemental meta information on several variables was added

CohortProfile:
        - further smoothing on ID_i has been done (should be stable... by now)

MethodsCAWI:
        - variables added: tx80102, tx80103, tx80210, tx80310
        - erroneously added waves 6 and 8 were dropped

pTargetCATI:
        - versionized variables had to be added due to change of item-battery-
            composition:
                tg51101_v1, tg51102_v1, tg51103_v1, tg51104_v1, tg51108_v1, tg51109_v1,
                     tg51110_v1, tg51111_v1, tg51112_v1, tg51113_v1, tg51114_v1,
                    tg51115_v1, tg51116_v1, tg51117_v1, tg51118_v1
        - polarity of categories in variable t428050 have changed since wave 11 as the
            field instrument was edited in this way:
                - in releases prior to version 11 the coding was: 1="not at all"; 2="
                    hardly"; 3="average"; 4="strongly"; 5="very strongly";
                - since version 11 the coding is: 1="very stringly"; 2="strongly"; 3="
                    average"; 4="hardly"; 5="not at all"

spPartner:
        - variable ts31410 corrected

xEcoCAPI:
        - plausible values for competency data were added

xPlausibleValues:
        - new dataset since release 13-0-0: provides plausible values for competency
            data stored in xTargetCompetencies



=====================================================
* Changes introduced to NEPS:SC5 by version 13.0.0 *
=====================================================

*******************************************
Known issues:
MethodsCAWI: waves 6 and 8 erroneously added to MethodsCAWI, data in those lines is
    completely missing, please drop these waves

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


General remarks:
        - some versionized variables were dropped, some were introduced ...
        - some intro variables are back again ...
        - some variable labels were corrected
        - supplemental meta information on several variables was added

CohortProfile:
        - some checks on plausibility and smoothing on ID_i has been done
```

```
EditionsBackup:
        - new dataset since release 12-0-0: provides original data prior to coding and
            smoothing during the process of data preparation

pTargetCATI:
        - variable tg26390_g1 "Spell number with reference to transition questions (
            from spEmp)" was generated for merging information from spEmp to
            information on transitions into the labormarket in pTargetCATI

spVocTrain:
        - variable t724401 (grades of academic degrees) dropped - information is
            integrated into variable ts15265 (the variable concerning grades of
            vocational qualification)
        - variable ts15219_g1 dropped - information provided in variable ts15219_g1 is
            redundant to information provided by variable ts15219


=====================================================
* Changes introduced to NEPS:SC5 by version 12.0.0 *
=====================================================


General remarks:
        - for several variables information of the respective _v-variables was
            integrated into the variables without suffixes. The respective _v-
            variables were dropped.
        - intro-variables were droppped, except for intro-variables in spChild and
            spPartner.

pTargetCATI:
        - variable tg24201_g1, tg24202_g2 and tg02001_ha were generated to provide
            detailed information on teaching degrees gathered in wave 1. For further
            information, see Data Manual (5.4 Teacher Education Students and Teachers)
        - variable tg12001_g2 was generated to provide missing information on the
            desired subject of study for target persons who claim to study in their
            desired subject. Therefore it combines information from variable tg04001
            and tg12003.

pTargetCAWI:
        - for several variables open answers were (belatedly) coded.

spChild:
        - variable ts33204_g1 was generated to provide information on the status of the
             child. Therefore category "other child in household" was added.

spEmp:
        - variable tg2608a "student or other occupation" has been renamed to ts23256 to
             match the corresponding variable's name in other starting cohorts.

spSchoolExtExam:
        - additional information on external examinations from wave 1 and 3 was
            gathered from file spSchool.

spVocExtExam:
        - additional information on external examinations from wave 1 and 3 was
            gathered from file spVocTrain.

spVocPrep:
        - variable ts13101 was deleted by mistake. Please use information on the
            program type for wave 1 and 3 from earlier SUF releases.
```

```
spVocTrain:
        - variables tg24162_g1, tg24165_g1 and tg24168_g1 were generated to provide
            information on major or minor subjects for each subspell of an episode.For
             further information, see Data Manual (5.1 service variables).
        - information on external examinations from waves 1 and 3 was removed and
            integrated in file spVocExtExam.
        - variable ts15221_g1 was edited to provide (the revised) information on the
            intended vocational qualification for all target persons and for all
            subspells of an episode. For further information, see Data Manual (5.1
            service variables).
        - variable tg01003_ha was edited and now excludes administration and business
            academies.
        - servicevariables with information on subject of studies (tg2417*) were
            revised.


====================================================
* Changes introduced to NEPS:SC5 by version 11.0.0 *
====================================================


General remarks:
        - several variables surveyed have been renamed to *_v1 and *_v2 in prior
            releases;
                this has been improved by renaming some variables with suffix _v1 to
                    variable names without suffixes
                and some variables with suffix _v2 to suffix _v1;
                a detailed list and comparison of _v1 variables can be found in the
                    Data Manual (Appendix A.3).

CohortProfile:
        - testy testm testd erroneously coded to -56 for testing data in wave 7 have
            now been coded with correct dates

pTargetCAWI:
        - there have been changes during the field phase regarding interviewer
            instructions in variable "tg51001";
                the new indicator variable "Version_tg51001" contains information about
                     the version of the survey instrument

MethodsCAWI:
        - a new data file including para data from the CAWI interviews has been added


====================================================
* Changes introduced to NEPS:SC5 by version 10.0.0 *
====================================================

General remarks:
        - several variables surveyed prior to wave 10 have been renamed to *_v1 and *
            _v2,
                as wording of question texts has changed in recent survey instruments

CohortProfile:
        - testy testm testd erroneously had been coded to -56 even though tx80522==1;
            this has been fixed
        - new indicator variable tx80121 has been introduced: subsample "students of
            economics"
        - tx80921 has been revised

xEcoCAPI:
```

```
        - new dataset featuring items from CAPI-shortquestionaire, economics-competency
            -test and the corresponding
                methods data that has been administered to students of economics in
                    wave 7; all of these data
                has been removed from pTargetCATI, xTargetcompetencies, and
                    MethodsCompetencies, respectively, for this subsample


==================================================
* Changes introduced to NEPS:SC5 by version 9.0.0 *
==================================================

pTargetCATI:
        - ts15911 (highest degree obtained) was falsely programmed in wave 9. Therefore
             ts15911_g1 was generated for all participants.

spVocTrain:
        - original variables tg2416* (subjects) were edited due to discrepancies
            between subspells. Subsequently, subjects are filled for the first
            explicit mention only. Missing information was labeled accordingly.
            Working with service variables is recommended.
        - service variables tg2417* (subjects) have been revised so that each subspell
            of a corresponding spell is now filled with the first information
            available, still variables tg24170_g1-_g5, tg24173_g1-_g5 and tg24176_g1-
            _g5 provide complete information for all study episodes.
        - ts15221 (qualification sought) was falsely derived in some cases. Therefore,
            ts15221_g1 was generated for the affected episodes



==================================================
* Changes introduced to NEPS:SC5 by version 8.0.0 *
==================================================

General remarks on harmonization of variables concering subjects, type of university
    and type of vocational training program:
        - harmonization of type of university-variable:  tg01003_g1(pTargetCATI) >>
            tg01003_ha (spVocTrain, considering values of ts15201)
        - harmonized service variables on subjects: tg24160_g*, tg24163_g*, tg24166_g*
            (spVocTrain) >>  tg24170_g*, tg24173_g*, tg24176_g* in spVocTrain (
            considering values of tg04001_g1-5, tg04004_g1-5, tg04007_g1-5 in
            pTargetCATI)
        - harmonization provides valid values for type of university and subjects where
             information on study episode from winter term 2010/11 was missing
        - missing codes -28, -29 were introduced in the original variables tg24160_g*,
            tg24163_g*, tg24166_g*, tg01003_g1, ts15201


CohortProfile:
        - tx80951 indicates the participation status for students of economics in wave
            7. Besides CATI survey and competency testing, these students had also the
             possibility of taking parting in a short CAPI questionaire as well.


pTargetCATI:
        - 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 pTargetCATI dataset have been renamed using the "v1"
            suffix [t400500_g1v1,t400500_g2v1,t400500_g3v1], and the new ones have
            been introduced
```

```
        - variables of students of economics who took part in a short CAPI questionaire
            were added to pTargetCATI


spVocTrain:
        -  service variables tg2417* (subjects) and tg01003_ha  (type of university)*
            were introduced to simplify working with the dataset. Small discrepancies
            from the original variables (tg2416*) cannot be ruled out and have to be
            considered by the user.
        - each subspell of a corresponding spell was filled with the most recent
            information available, so that the variables tg24170_g1-5, tg24173_g1-5,
            tg24176_g1-5 provide complete information for all study episodes.


==================================================
* Changes introduced to NEPS:SC5 by version 6.0.0 *
==================================================

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 waves 5 (CAWI) and 6 (CATI/CAPI) have been incorporated into the
            data
        - the subspell harmonization routine in all spell datasets ("sp*") has been
            updated, leading to more accurate harmonized subspell information (
            subspell==0) for panel continuation spells
        - staff from NEPS stage 7 at the DZHW excessively reviewed and overworked all
            syntax for generated tg*-variables, which may lead to slightly different
            contents
        - staff from NEPS stage 7 at the DZHW reviewed the cohorts' sample frame in
            consultation with NEPS methods department, leading to 3 observations
            removed from the SUF
        - all datasets from version 4.0.0 did not reflect the correct doi in their
            dataset labels; the correct doi would have been "10.5157/NEPS:SC5:4.0.0",
            not "none";
                this issue has been fixed and all datasets of version 6.0.0 correctly
                    are labeled with doi:10.5157/NEPS:SC5:6.0.0

xTargetCompetencies:

        - all variables of domains "maths" and "reading" erroneously contained the
            missing value -54 ("missing by design") in versions 4.0.0 and 3.1.0;
                as there were no additional competency assessments in wave 4, it was
                    safe to use the xTargetCompetencies dataset file from version
                    3.0.0
                instead without missing any information; this has been fixed

pTargetCATI:

        - variables "Specialized fair/congress: professional/personal reasons" [
            t272802_w1] and "Specialized fair/congress: Learned something new" [
            t272802_w1]
                as well as the corresponding variables for "Lectures" [t272802_w2,
                    t272802_w2] and "Self-instruction programs" [t272802_w3,t272802_w3
                    ] in version 4.0.0 and earlier
```

```
                erroneously are not filled for all interviewees reporting the specific
                    further education activity; this has been fixed
        - variable names of variables "Father's mother: Country of birth" [t405240*]
            and "Mother's father: Country of birth" [t405230*] in dataset pTargetCATI
                erronneously had been flipped in version 4.0.0, also leading to slight
                    inconsistencies in generated variables for migrational background;
                     this has been fixed

spChild:

        - all wide variables documenting cohabitation (*_w*) in version 4.0.0 and
            earlier with the focal child have been extracted and are now saved in the
            separate dataset "spChildCohab"

spChildCohab:

        - new dataset containing chidl cohabitation spells that formerly had been saved
             in wide format inside of spChild

spEmp:

        - version 4.0.0 and earlier did not contain coded occupational information for
            studentical employment episodes reported in wave 1; this has been fixed

Biography:

        - additional spells of type "data edition gap" have been inserted to fill gaps
            between
                (a) the eighth birth day and the first reported episode and
                (b) the most recently reported episode and the most recent interview
                    date

==================================================
* Changes introduced to NEPS:SC5 by version 4.0.0 *
==================================================

General:

        - full translations have been added
        - wave 4 (online survey in semester 5) has been added
        - several minor bug fixes to data edition scripts have been introduced

pTargetCATI:

        - when generating variable "Global self-esteem" [t66003a_g1] in the pTargetCATI
             dataset, variable "Global self-esteem: competence" [t66003d] erroneously
            had been ignored;
                this has been fixed;
                t66003a_g1 can be re-generated in 3.1.0 using the following Stata
                    syntax:
                * -------------------------BEGIN Stata-------------------------------
                local target_variable t66003a_g1
                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--------------------------------

xTargetCAWI:

        - as wave 3 data makes this a panel dataset, the filename has changed from "
            xTargetCAWI" to "pTargetCAWI"


==================================================
* Changes introduced to NEPS:SC5 by version 3.1.0 *
==================================================

General:
        - meta data in all datasets have been revised and updated where appropriate
        - English translation for all datasets except xTargetCAWI have been introduced
            to the data
        - end dates in episodes neglected in the panel interview erroneously contained
            the interview
                date of the panel wave instead of the first interview's date; this has
                    been fixed
        - 185 duplicate respondents have been identified by the survey institute;
                the redundant observations have been dropped from the data, resulting
                    in slightly smaller number of cases

pTargetCATI:
        - variables indicating migrational background (t400500_g1 through _g3) have
            been added

spVocTrain:
        - spell integration and recommendation (via variable tx20100) was erroneous;
            this has been fixed
        - spell linkage between waves 1 and 3 was erroneous; this has been fixed

spEmp:
        - spell linkage between waves 1 and 3 was erroneous; this has been fixed

Weights:
        - dataset containing weighting variables has been added

Basics:
        - dataset containing oversimplified, "flat" cross-sectional data on the cohort
            has been added;
                use for orientation, not for analyses!

xInstitution:
        - dataset containing detailed information on the targets' institutions has been
             added for onsite access in Bamberg
```

B.3 Comparison of _v1 variables

The following tables shows all changes of variables where construction of a `_v1`-variable seemed necessary. Note that by v1, we generally mean *first version* or *version one*. Thus, this usually is the old variant of a variable, which has been updated in a later wave. Small arrows indicate if an entry belongs to the old version («) or if it is an update (»). Grayed out entries did not change between the versions, and are printed for your orientation only.

`pTargetCATI`

	« t400500_g1v1 | pTargetCATI | t400500_g1 »	
Label		Generation status
Text		
-54		missing by design
0		no migrant background
1		1st generation
2		1.5th generation
3		2nd generation
4		2.25th generation
5		2.5th generation
6		2.75th generation
7		3rd generation
8		3.25th generation
9		3.5th generation
10	»	3.75th generation

	« t400500_g2v1 | pTargetCATI | t400500_g2 »	
Label		Missing/contradicting imformation about country of birth for generation status
Text		
-54		missing by design
1		Unique assignment possible
2		Information for target person unknown
3		Information for one parent unknown
4		Information for both parents unknown
5		Information for one grandparent unknown
6		Information for two grandparents unknown
7		Information for three grandparents unknown
8		Information for four grandparents unknown
9		no assignment to a generation status possible

	« t400500_g3v1 | pTargetCATI | t400500_g3 »
Label	Group of origin
Text	
-54	missing by design
1	Germany
2	Italy
3	Poland
4	Romania
5	Turkey
6	Former Yugoslavia
7	Former Soviet Union
8	Central and South Amerika, Carribean
9	Northern and Western eurospe
10	North America
11	Oceania/Polynesia
12	other country of Middle East and North Africa
13	other country of Africa
14	other country of Asia
15	other Central and Eastern Europe
16	other Southern Europe
17	abroad, but cannot be assigned to a specific group of origin

	« t514008_v1 | pTargetCATI | t514008 »	
Label	«	Satisfaction with higher education
	»	Satisfaction with course of study
Text		How satisfied are you with your higher education?
-98		don't know
-97		refused
-93		does not apply
-54		missing by design
0		completely dissatisfied
1		1
2		2
3		3
4		4
5		5
6		6
7		7
8		8
9		9
10		completely satisfied

	«	**t516201_v1 | pTargetCATI | t516201** »
Label	«	Party election
	»	Parliamentary elections: Party election
Text		If parliamentary elections were to be held tomorrow, which party would you give your second vote to?
-98		don't know
-97		refused
-93	»	does not apply
-55	«	not determinable
-54		missing by design
-21	»	would not vote
-20		not entitled to vote, because no German citizenship
1		CDU or CSU
2		SPD - Social Democratic Party of Germany
3	«	FDP (political party)
	»	FDP - Free Democratic Party
4		Bündnis 90/Die Grünen [green political party]
5		Die Linke - Left Party
6		NPD - National Democratic Party of Germany
7	«	Die Republikaner - The Republicans
8		other party
9		Would not vote
10		Piratenpartei - Pirate Party
11	»	AfD

	«	t516202_g1v1 | pTargetCATI | t516202_g1 »
Label		Party election (another party)
Text		Which other party is this?
-98		don't know
-97		refused
-55		not determinable
-54		missing by design
-52		implausible value removed
1		Citizens' initiatives
2		Voter participation with invalid vote
3		Indifferent
4		Already disbanded parties (graue Panther 2008, Deutsche Bierunion after 1990, SDP after 1990)
5		APPD - Anarchistic pogo party Germany
6	«	AUF-Party, for work, environment, family (Christain)
	»	AUF-party, for work, environment, family (Christian)
7		Bayernpartei - Bavarian party
8		PBC - Party of Bible-abiding Christians
9		BIG - Alliance for innovation and justice (party of Germans of Turkish origin)
10		Die Frauen - Feminist party The Women
11		Die Freien Wähler - The Free Voters
12		Die Freiheit - Civil rights' party for more freedom and democracy
13		Die PARTEI - Party for Labour, Rule of Law, Animal Protection, Promotion of Elites and Grassroots Democratic Initiative
14		Die Tierschutz-Partei - Party Humans Environment Animal Rights
15		Die Violetten - The Purples
16		DKP - German Communist Party
17		FAMILIE - Family Party of Germany
18	«	MLPD - Marxist-Leninist Party of Germany
	»	MLPD - Marxist–Leninist Party of Germany
19		ÖDP - Ecological Democratic Party
20		PDV - Party of Reason
21	«	Pro NRW - Pro North-Rhine Westphalia
	»	Pro NRW
22		SSW - South Schleswig Voters' Association
23	«	AfD - Alternative for Germany
24	«	Bündnis 21/RRP - Pensioners' Party
	»	Bündnis 21/RRP - Rentnerinnen- und Rentner-Partei [Pensioners' Party]
25		KPD - Communist Party of Germany
26		UDP - Union of German Patriots
27	»	Alfa (Allianz für Fortschritt und Aufbruch [alliance for progress and awakening]) since 2016 LKR (Liberal-Konservative Reformer [liberal-conservative reformers])
28	»	Deutsche Mitte - German Middle
29	»	V-Partei (Party for Change, Vegetarians and Vegans)
30	»	Liberale (New Liberals - The Social Liberal)
31	»	Die Republikaner - The Republicans

	«	**t525008_v1 | pTargetCATI | t525008** »
Label		Smoking status
Text	«	Did you smoke in the past or do you currently smoke?
	»	Do you currently smoke - even if only occasionally?
-98	«	don't know
-97	«	refused
-54		missing by design
1	«	have never smoked
	»	yes, daily
2	«	did smoke before
	»	yes, occasionally
3	«	currently smoke occasionally
	»	no, not anymore
4	«	currently smoke every day
	»	have never smoked

	«	**t525209_v1 | pTargetCATI | t525209** »
Label	«	Alcohol consumption
	»	Alcohol consumption frequency last 12 months
Text	«	How often do you consume alcoholic drinks?
	»	How often do you consume alcoholic drinks? Think about the average over the last 12 months.
-98	«	don't know
-97		refused
-54		missing by design
1	«	(almost) never
	»	never
2		once a month or less
3		twice or three times a month
4		once a week
5		several times a week
6	«	(almost) every day
	»	daily

	«	**tg2450a_v1 | pTargetCATI | tg2450a** »
Label	«	Doctorate context - research project higher education institution
	»	Doctorate context - third-party funded position higher education institution
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

	«	**tg2450b_v1 | pTargetCATI | tg2450b** »
Label	«	Doctorate context - chair higher education institution
	»	Doctorate context - budget position higher education institution
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

	«	**tg2450c_v1 | pTargetCATI | tg2450c** »
Label		Doctorate context - non-university research institution
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

	«	**tg2450d_v1 | pTargetCATI | tg2450d** »
Label		Doctorate context - doctoral program
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

	«	**tg2450e_v1 | pTargetCATI | tg2450e** »
Label	«	Doctorate context - doctorate course of study
	»	Doctorate context - scholarship program
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

	«	**tg2450f_v1 | pTargetCATI | tg2450f** »
Label	«	Doctorate context - private sector/industry
	»	Doctorate context - private sector (industrial research and development)
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

		« tg2450g_v1 | pTargetCATI | tg2450g »
Label		Doctorate context - alongside studies
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

		« tg2450h_v1 | pTargetCATI | tg2450h »
Label	«	Doctorate context - without institutional integration
	»	Doctorate context - without institutional integration, free doctorate student
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

		« tg2450i_v1 | pTargetCATI | tg2450i »
Label		Doctorate context - other
Text	«	[MF] We have noted that you have begun a doctorate. In the following, we would like to ask you a few questions about this doctorate. Under what circumstances are you currently studying for a doctorate?
	»	We have noted that you have started your doctoral program. In the following we would like to ask you some questions. A doctoral thesis can be written in very different institutional contexts, e.g. at a higher education institution or a research institution as a research assistant, in a structured doctoral program or as a free doctoral student without institutional integration. In which institutional context are you currently doing your doctorate?
-98		don't know
-97	«	refused
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
-20	»	none of it
0		not specified
1		specified

		« tg60013_v1 | pTargetCATI | tg60013 »
Label		Auxiliary variable: phase of teacher education and employment (CATI)
Text	«	[AUTO] Auxiliary variable: teaching groups, current status
	»	[AUX] Auxiliary variable: Teaching groups, current status
-92	»	question erroneously not asked
-54		missing by design
0		no teaching reference or status unknown
1		first phase teacher training not yet completed
2	«	completed teaching degree course and Referendariat is intended or completed teaching degree course and employment as a teacher is intended
	»	completed teaching degree course and intended Referendariat [period as a trainee teacher] or completed teaching degree course and intended employment as a teacher
3		ongoing teaching Referendariat
4		completed Referendariat [period as a trainee teacher] and employment as a teacher is intended
5		employment as a teacher
6	»	interrupted employment as a teacher (e.g. due to parental leave)

	«	**tg60022_v1 | pTargetCATI | tg60022** »
Label	«	Screening employed teachers
	»	Screening Employed teachers
Text	«	As announced at the beginning of the interview, there is a survey section for student teachers, trainee teachers and teaching staff Just to be on the safe side and to simplify the interview process, would you please tell me briefly whether you are currently employed as a teacher?
	»	As announced at the beginning of the interview, the interview includes a survey section for student teachers, Referendare [trainee teachers] and teachers. Just to make sure and to simplify the interview process, please tell me briefly if you are currently employed as a teacher.
-98	«	don't know
-97	«	refused
-93	«	does not apply
-92	«	question erroneously not asked
-54		missing by design
-52	«	implausible value removed
1		yes
2		no
3	»	yes, but I have currently interrupted my employment as a teacher

	«	**tg60031_v1 | pTargetCATI | tg60031** »
Label	«	Preload completed teaching degree course
	»	Preload completed teacher education program (CATI), as of 13th wave
Text		[AUTO] Preload Completed teaching degree course
-54		missing by design
0		no teaching degree course completed
1		teaching degree course completed

	«	**th21300_v1 | pTargetCATI | th21300** »
Label		Number of selected courses
Text	«	[AUTO]: Via random selection, select two courses that were completed between <intmPRE/intmjPRE> and <20102(intm/intj)>
	»	[AUTO]: Randomly select one of the courses that was completed from <intmPRE/intmjPRE> to <20102(intm/intj)>
-54		missing by design
0		no course selected
1		1 course selected
2		2 courses selected

	«	**ts15911_v1 | pTargetCATI | ts15911** »
Label	«	Graduate
	»	Auxiliary variable: highest degree
Text	«	[AUX]
	»	[AUX] Highest degree
-54		missing by design
0		no higher education qualification
1	«	BA, MA, Diploma, state examination
	»	BA
2	«	Doctorate
	»	MA, Diploma, state examination
3	»	Doctorate

pTargetCAWI

	«	**t242400_g2v1 | pTargetCAWI | t242400_g2** »
Label	«	AUX: subject group ref subject questions about learn. env. (destatis 2010/11)
	»	Subject group reference subject learning environment (destatis 2010/11)
Text	«	[AUTO] Auxiliary variable: Field of study referenced for questions about learning env.
	»	The following is about your experiences in your current course of study. If you are studying several subjects, they may be very different, e.g. in terms of content and/or organization of teaching. Therefore, please select the major or teaching subject to which you would like to refer your information in the next questions.
-99		filtered
-97		refused
-96		not in list
-92	«	question erroneously not asked
-91		survey aborted
-55	»	not determinable
-54		missing by design
-29		value from the last sub-episode
-28		value from recruitment pTargetCATI
-20		no further subject
1		Linguistic and cultural studies
2		Sport
3		Law, economics and social science
4		Mathematics, sciences
5		Human medicine/health sciences
6		Veterinary medicine
7		Agricultural-, forest- and nutrition sciences
8		engineering
9		Arts, art science
10		Outside the study area structure

	«	**t242400_g5v1 | pTargetCAWI | t242400_g5** »
Label	«	AUX: ISCED-97 ref subject questions about learning environment (1-digit level)
	»	ISCED-97 reference subject learning environment (1-digit level)
Text	«	[AUTO] Auxiliary variable: Field of study referenced for questions about learning env.
	»	The following is about your experiences in your current course of study. If you are studying several subjects, they may be very different, e.g. in terms of content and/or organization of teaching. Therefore, please select the major or teaching subject to which you would like to refer your information in the next questions.
-99		filtered
-98		don't know
-97		refused
-96		not in list
-92		question erroneously not asked
-91		survey aborted
-55		not determinable
-54		missing by design
0		general educational programs
1		Education
2		Humanities and Arts
3		Social sciences, Business and Law
4		Natural sciences, mathematics and computer science
5		engineering, manufacturing and construction
6		Agriculture and Veterinary
7		Health and welfare
8		Services
9		not known or unspecified

	« t289900_v1 | pTargetCAWI | t289900 »	
Label		Type of accommodation
Text		Now we would like to ask you a few questions about your living situation and your spending. During term time, do you stay primarily...
-99	«	filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
1	«	with parents or relatives?
	»	with your parents?
2		in a dormitory?
3	«	in some other rental accommodation?
	»	in another type of rented apartment?/in a rented apartment?
4	«	in an apartment/house that you own?
	»	in a condo/own house?
5		with private individuals for subtenancy?
6	»	with relatives?

	« tg51101_v1 | pTargetCAWI | tg51101 »	
Label		Current activity: employed
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently …
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

		« tg51102_v1 | pTargetCAWI | tg51102 »
Label		Current activity: Volontariat
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

		« tg51103_v1 | pTargetCAWI | tg51103 »
Label		Current activity: internship
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

		« tg51104_v1 | pTargetCAWI | tg51104 »
Label	«	Current activity: vocational training
	»	Voc. train./further educ.: vocational training
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you currently ...?
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-21	»	none of both
0		not specified
1		specified

	«	**tg51108_v1 | pTargetCAWI | tg51108** »
Label	«	Current activity: retraining or further education
	»	Voc. train./further educ.: retraining, further education
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you currently ...?
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-21	»	none of both
0		not specified
1		specified

	«	**tg51109_v1 | pTargetCAWI | tg51109** »
Label	«	Current activity: (voluntary) services, (military/alternative/community/social)
	»	Other activities: volunt. military service/social year/fed. volunt. service
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51110_v1 | pTargetCAWI | tg51110** »
Label	«	Current activity: on parental leave
	»	Other activities: parental leave
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51111_v1 | pTargetCAWI | tg51111** »
Label	«	Current activity: housewife/househusband
	»	Other activities: housewife/househusband
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51112_v1 | pTargetCAWI | tg51112** »
Label	«	Current activity: unemployed
	»	Other activities: unemployed
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51113_v1 | pTargetCAWI | tg51113** »
Label	«	Current activity: on sick leave
	»	Other activities: ill
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51114_v1 | pTargetCAWI | tg51114** »
Label	«	Current activity: other
	»	Other activities: other, namely:
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	Are you also or exclusively doing any of the following activities? I am currently ...
-99	«	filtered
-97		refused
-92	«	question erroneously not asked
-91		survey aborted
-54		missing by design
-20	»	none of it
0		not specified
1		specified

	«	**tg51115_v1 | pTargetCAWI | tg51115** »
Label		Current activity: Referendariat
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

	«	**tg51116_v1 | pTargetCAWI | tg51116** »
Label		Current activity: vicariate
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

	« tg51117_v1 | pTargetCAWI | tg51117 »	
Label		Current activity: trainee program
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

	« tg51118_v1 | pTargetCAWI | tg51118 »	
Label		Current activity: probationary year / practical year
Text	«	[MF] Which of the following activities are your currently doing? I am currently ...
	»	[MF] Which of the following positions are you currently working in? I am currently ...
-99		filtered
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
0		not specified
1		specified

	«	**tg51300_v1 | pTargetCAWI | tg51300** »
Label	«	Change of subject since starting university
	»	Change of subject since last survey
Text	«	Have you changed your field of study since starting your studies in winter semester 2010/2011?
	»	Have you changed your subject of study since <h_zebePRE(label)>?
-99		filtered
-98		don't know
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
-52		implausible value removed
1		yes
2		no

	«	**tg51400_v1 | pTargetCAWI | tg51400** »
Label	«	Change in leaving qualification since starting university
	»	Change Leaving qualification since last survey
Text	«	Have you switched your chosen leaving qualification since the starting your studies in winter semester 2010/2011 (for example, from a bachelor's degree to a state examination)?
	»	Have you changed the leaving qualification since <h_zebePRE(Label)> (for example, from a Bachelor degree to a state examination)?
-99		filtered
-98		don't know
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
-52		implausible value removed
1		yes
2		no

	«	**tg51500_v1 | pTargetCAWI | tg51500** »
Label	«	Change in university after starting studies
	»	Change of higher education institution since last survey
Text	«	Have you changed universities since starting your studies in winter semester 2010/2011?
	»	Have you changed higher education institution since <h_zebePRE(Label)>?
-99		filtered
-98		don't know
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
-52		implausible value removed
1		yes
2		no

	«	**tg60012_v1 | pTargetCAWI | tg60012** »
Label	«	Auxiliary variable: phase of teacher education and employment (CATI)
	»	Auxiliary variable: phase of teacher training and employment (CAWI)
Text	«	[AUTO] Auxiliary variable: teaching groups, current status
	»	[AUTO] Auxiliary variable: phase teacher training and employment (CAWI)
-54		missing by design
0		no teaching reference or status unknown
1		first phase teacher training not yet completed
2	«	completed teaching degree course and Referendariat is intended or completed teaching degree course and employment as a teacher is intended
	»	completed teaching degree course and intended Referendariat [period as a trainee teacher] or completed teaching degree course and intended employment as a teacher
3		ongoing teaching Referendariat
4		completed Referendariat [period as a trainee teacher] and employment as a teacher is intended
5		employment as a teacher
6	»	interrupted employment as a teacher (e.g. due to parental leave)

	«	**tg60017_v1 | pTargetCAWI | tg60017** »
Label		Auxiliary variable: phase teacher education and employment (CAWI), update
Text	«	[AUTO] Auxiliary variable: teaching groups, status (updated)
	»	[AUTO] Auxiliary variable: phase teacher training and employment (CAWI), update
-54		missing by design
0		no teaching reference or status unknown
1		first phase teacher training not yet completed
2	«	completed teaching degree course and Referendariat is intended or completed teaching degree course and employment as a teacher is intended
	»	completed teaching degree course and intended Referendariat [period as a trainee teacher] or completed teaching degree course and intended employment as a teacher
3		ongoing teaching Referendariat
4		completed Referendariat [period as a trainee teacher] and employment as a teacher is intended
5		employment as a teacher
6	»	interrupted employment as a teacher (e.g. due to parental leave)

	«	**tg60025_v1 | pTargetCAWI | tg60025** »
Label		Teaching: current or intention
Text		Are you currently employed as a teacher or would you like to become a teacher?
-99		filtered
-91		survey aborted
-54		missing by design
1	«	yes, I am a teacher
	»	yes, I work as a teacher
2		yes, I would like to become a teacher
3		no, neither
4	»	yes, but I have currently interrupted my employment as a teacher (e.g. due to parental leave)

	«	**tg71121_v1 | pTargetCAWI | tg71121** »
Label	«	Doctorate subject_Update_1
	»	Doctorate subject - update (yes/no)
Text	«	In your last telephone interview, you stated that you are doing your doctorate in <tg70002(Label)>. Is that still correct?
	»	Are you still doing your doctorate in the subject you told us about in the last phone interview in <h_zebePRE(Label)>?
-99		filtered
-98		don't know
-97		refused
-92		question erroneously not asked
-91		survey aborted
-54		missing by design
-52		implausible value removed
1		yes
2		no

spEmp

	«	**ts23228_v1 | spEmp | ts23228** »
Label	«	Type of education required
	»	Type of required training
Text		What kind of training is usually required to do this job?
-98		don't know
-97		refused
-92	«	question erroneously not asked
-55		not determinable
-54		missing by design
1		no qualification
2		a training on the job
3		a completed vocational training
4		a completed training at a Fachschule
5		a master craftsman's/craftswoman's certificate or technician certificate
6	«	a completed higher education qualification (university of applied sciences or university)
7		a doctorate or habilitation
8	»	a Bachelor's degree (university of applied sciences or university)
9	»	a Master's degree or state examination, a diploma or a Magister (degrees from a university of applied sciences or university)

	« ts23901_v1 | spEmp | ts23901 »	
Label		Auxiliary variable: current employment
Text		[AUX] Auxiliary variable Current employment
-95	«	implausible value
-55		not determinable
-54		missing by design
1	«	currently employed
	»	Current employment
2	«	employed within the last year, but not currently
	»	Completed employment
3	«	not employed within the last year / end not assignable

	« ts23911_v1 | spEmp | ts23911 »	
Label		Auxiliary variable: type of employee
Text		[AUX] Employee type
-55		not determinable
-54		missing by design
-29	»	value from the last sub-episode
-20	«	not assignable
1	«	Worker/ employee
	»	worker/employee/civil servant/soldier/not classifiable
2	«	Civil servants/soldiers
	»	temporary/seasonal worker
3	»	2nd job market/training opportunities
4	»	self-employed/assistant/ freelancer
5	«	2nd job market
6	«	Freelancer
7	«	Self-employed
8	«	Positions in an assisting capacity
9	«	Vocational training positions
13	»	semi-skilled or unskilled work/student assistant
14	»	private student tuition/homework supervision

spInternship

	« tg36111_v1 | spInternship | tg36111 »	
Label		Average working hours Internship
Text		How many hours per week are your average working hours in this internship?
-98		don’t know
-97		refused
-55		not determinable
-54		missing by design
-21		no fixed working hours
-20		more than 50 hours per week

spPartner

	« ts31223_v1 | spPartner | ts31223 »	
Label		Employment partner
Text	«	Is your partner currently full-time employed, part-time employed or unemployed?
	»	Is your partner currently employed full-time or part-time, has a side-job or is unemployed?
-98		don’t know
-97		refused
-55		not determinable
-54		missing by design
1	«	primarily working
	»	full-time employed
2	»	part-time employed
3	«	part-time employed
	»	in a side job
4		unemployed

	«	**ts31510_v1 | spPartner | ts31510** »
Label	«	Termination of partnership (separation/death, moving out without separation)
	»	End of the partnership due to separation from or death of the partner
Text	«	Have you divorced, separated or is your partner deceased?
	»	Did you get divorced, did you split up, or did your partner die?
-98	»	don't know
-97		refused
-55		not determinable
-54		missing by design
1		divorced/civil partnership annulled
2		separated
3	«	Partner deceased
	»	partner deceased
4	»	marital status unchanged
5	»	moved back together, currently living together
6	»	living apart, but still in partnership
9	«	Do not live together any more, with partnership still persisting

spVocExtExam

« ts15304_v1 | spVocExtExam | ts15304 »

Label		External examination qualification
Text		What leaving qualification did you obtain?
-99	«	filtered
-98		don't know
-55	«	not determinable
-20		no qualification
1		completed apprenticeship (administrative, company-based, industrial, agricultural), journeyman's/journeywoman's certificate or apprenticeship certificate (craft certificate), dual vocational training
2		graduation from a school of public health
3		certificate from a Berufsfachschule [vocational school] or a Handelsschule [type of vocational school for commercial professions]
4		certificate from another Fachschule
5	«	master craftsman's/craftswoman's diploma
	»	Master craftsman's/craftswoman's diploma
6		technician certificate
7	«	diploma
8	«	Bachelor
9	«	Master
10	«	diploma from university of applied sciences (Dipl(FH))
	»	diploma from a university of applied sciences (Dipl(FH))
11	«	Diploma from a university
	»	diploma from a university
12		Bachelor (in teaching)
13		Bachelor (not in teaching)
14		Master (in teaching)
15		Master (not in teaching)
16		Magister
17		first state examination (in teaching)
18		first state examination (not in teaching)
19	«	second/third state examination
	»	second/third state examination (not in teaching)
20	«	Doctorate
	»	doctorate
21	«	Habilitation
	»	habilitation
22		medical specialist
23		civil service examination for the subclerical class
24		civil service examination for the clerical class
25		civil service examination for the executive class
26		civil service examination for the administrative class
27		IHK (Chamber of Industry and Commerce) examination
28		other qualification
29		other higher education qualification (e.g. ecclesiastical examination, artistic examination)
30	»	second state examination (in teaching)

spVocTrain

	« tg24205_v1 | spVocTrain | tg24205 »
Label	Point of time decision for Master
Text	When did you make the decision for your Master's degree program?
-55	not determinable
-54	missing by design
1	before starting the previous higher education program
2	during the previous higher education program
3	after completion of the previous course of study

	« ts15216_v1 | spVocTrain | ts15216 »	
Label	«	Course or training course with a certificate of attendance
	»	Course/training with graduation or certificate of attendance
Text	«	Is/was it planned to complete the course or training course with a qualification or a certificate of attendance or is/was neither of the two planned?
	»	Is/was it planned to complete the course or training with a graduation or with a certificate of participation, with a recognized license, with another certificate or with none of these?
-98	«	don't know
-97	«	refused
-55		not determinable
-54		missing by design
1	«	Qualification
	»	qualification
2		certificate of attendance
3	«	none of both
4	»	recognized license
5	»	another certificate
6	»	none of it

	«	**ts31228_v1 | spPartner | ts31228** »
Label	«	Exact professional position partner
	»	Exact vocational position partner
Text	«	And what is your (male) partner's exact professional position there?
	»	And what is your partner's exact occupational status there?
-98		Don't know
-97		Refused
-54		Missing by design
10		Unskilled worker
11		Semi-skilled worker/partially skilled worker
12		Skilled worker, journeyperson [trained craftsperson]
13		Assistant foreman, group leader, Brigadier [former GDR: Leader of a work unit]
14		Master, construction foreman
20		Low-skill occupation, e.g. salesperson
21		Qualified occupation, e.g. office clerk, technical draftsman
22		Highly qualified occupation or leading position, e.g. engineer, research assistant, department manager
23		Occupation involving extensive management duties e.g., director, CEO, member of the executive board
24		Production or plant foreman
30		In sub-clerical class (up to and including 'Oberamtsmeister')
31		In the clerical class, from assistant to principal secretary or office inspector, inclusively
32		Executive class (from inspector to Amtsrat inclusive and/or Oberamtsrat as well as elementary, secondary or intermediate school teacher inclusive)
33		In the administrative class, including judge, e.g. teacher starting from level of Studienrat [junior position held by school teachers upon career entry], senior government official
40		Military team rank
41		Non-commissioned officer, e.g. staff sergeant, sergeant, master sergeant
42		Simple officer to captain (included)
43		Staff officers from major to general/admiral
51		Self-employed as an academic, self-employed professional, e.g. physician, lawyer, architect
52		Self-employed person in agriculture
53		Self-employed person in trade, commerce, industry, service; other self-employment or entrepreneurship

	« ts31230_v1 | spPartner | ts31230 »	
Label		Management position partner
Text	«	Does your partner have a leading position in his activity?
	»	Does your partner hold a management position?
-98		Don't know
-97		Refused
-54		Missing by design
1		Yes
2		No

	« ts31410_v1 | spPartner | ts31410 »	
Label	«	Marriage / registered civil partnership
	»	Marriage/ registered civil partnership
Text	«	Did you marry your partner (<28109>)?
	»	Have you married your partner or have you registered the civil partnership?
-98		Don't know
-97		Refused
-54		Missing by design
1		Yes
2		No

	«	**ts3141m_v1 | spPartner | ts3141m** »
Label	«	Date of marriage (month)
	»	Marriage date (month)
Text	«	When did you marry your partner <28109>?
	»	When did you marry or register your civil partnership?
-98		Don’t know
-97		Refused
-93		Does not apply
-54		Missing by design
1		January
2		February
3		March
4		April
5		May
6		June
7		July
8		August
9		September
10		October
11		November
12		December
21		Beginning of the year/winter
24		Spring/Easter
27		Mid-Year/Summer
30		Fall
32		End of year

	« ts3141y_v1 | spPartner | ts3141y »	
Label	«	Date of marriage (year)
	»	Marriage date (year)
Text	«	When did you marry your partner <28109>?
	»	When did you marry or register your civil partnership?
-99		Filtered
-98		Don't know
-97		Refused
-96		Not in list
-95		Implausible value
-94		Not reached
-93		Does not apply
-92		Question erroneously not asked
-91		Survey aborted
-90		Unspecific missing
-56		Not participated
-55		Not determinable
-54		Missing by design
-53		Anonymized
-52		Implausible value removed
-51		No estimate in check module

	« ts31510_v1 | spPartner | ts31510 »	
Label		End of the partnership due to separation or death of a partner
Text		Did you get divorced, did you separate or is your (male) partner deceased?
-98		Don't know
-97		Refused
-54		Missing by design
1		Divorced / civil partnership annulled
2		Separated
3		Partner deceased
4		Marital status unchanged
5		Moved back in with partner, currently living together
6		No longer living together but partnership still exists

	« ts3151m_v1 | spPartner | ts3151m »	
Label	«	Date of partner's death (month)
	»	Date of death Partner (month)
Text		When did your partner pass away?
-98		Don't know
-97		Refused
-93		Does not apply
-54		Missing by design
1		January
2		February
3		March
4		April
5		May
6		June
7		July
8		August
9		September
10		October
11		November
12		December
21		Beginning of the year/winter
24		Spring/Easter
27		Mid-Year/Summer
30		Fall
32		End of year

	« ts3151y_v1 | spPartner | ts3151y »	
Label	«	Date of partner's death (year)
	»	Date of death Partner (year)
Text		When did your partner pass away?
-99		Filtered
-98		Don't know
-97		Refused
-96		Not in list
-95		Implausible value
-94		Not reached
-93		Does not apply
-92		Question erroneously not asked
-91		Survey aborted
-90		Unspecific missing
-56		Not participated
-55		Not determinable
-54		Missing by design
-53		Anonymized
-52		Implausible value removed
-51		No estimate in check module

	« ts3152m_v1 | spPartner | ts3152m »	
Label		Date of moving apart (Month)
Text	«	When did you or your partner move out of the shared home?
	»	When did you or your partner moved out of the common household?
-98		Don't know
-97		Refused
-93		Does not apply
-54		Missing by design
1		January
2		February
3		March
4		April
5		May
6		June
7		July
8		August
9		September
10		October
11		November
12		December
21		Beginning of the year/winter
24		Spring/Easter
27		Mid-Year/Summer
30		Fall
32		End of year

	« ts3152y_v1 | spPartner | ts3152y »	
Label		Date of moving apart (Year)
Text	«	When did you or your partner move out of the shared home?
	»	When did you or your partner moved out of the common household?
-99		Filtered
-98		Don’t know
-97		Refused
-96		Not in list
-95		Implausible value
-94		Not reached
-93		Does not apply
-92		Question erroneously not asked
-91		Survey aborted
-90		Unspecific missing
-56		Not participated
-55		Not determinable
-54		Missing by design
-53		Anonymized
-52		Implausible value removed
-51		No estimate in check module

spVocExtExam

	« ts15304_v1 | spVocExtExam | ts15304 »	
Label		External examination qualification
Text		What leaving qualification did you obtain?
-20		no qualification
1		Completed apprenticeship (commercial, corporate, trade-oriented, agricultural) journey person's or assistant's certificate (skilled worker's certificate), dual vocational education and training
2		Leaving certificate from a school for health care professionals
3		Leaving certificate of Berufsfachschule, leaving certificate of a commercial school
4	«	Other type of leaving certificate of the Fachschule
	»	other type of leaving certificate from a Fachschule
5		Master's / foreman's certificate
6		Technician's certificate
10		Diplom from a university of applied sciences (Dipl(FH))
11		Diplom from a university
12		Bachelor's degree teaching profession
13		Bachelor (not for teaching post)
14		Master teaching post
15		Master (not for teaching post)
16	«	Magister
	»	Magister [German degree in tertiary education, pre-Bologna system, level equivalent to master]
17		First state examination for teaching post
18		First state examination (not for teaching post)
19	«	Second or third state examination
	»	Second/Third State Examination (without teaching post)
20		Doctorate
21		Habilitation
22		Medical specialist
23		Civil service examination for the subclerical class
24		Civil service examination for the clerical class
25		Civil service examination for the executive class
26		Civil service examination for the administrative class
27		IHK (Chamber of Industry and Commerce) examination
28		Other leaving qualification
29	«	Other degree from a higher education institute (e.g., ecclesiastical examination, artistic examination)
	»	Other degree from a higher education institution (e.g., ecclesiastical examination, artistic examination)
30	»	Second State Examination teaching post

spVocTrain

	« tg24146_v1 | spVocTrain | tg24146 »	
Label	«	Change of type of leaving qualification as against pre-episode
	»	Change of type of qualification compared with pre-episode
Text	«	Will your next degree course result in the same leaving qualification as the degree course we talked about before, or is it another leaving qualification, e.g. Bachelor instead of state examination or elementary school teaching qualification instead of Gymnasium teaching qualification?
	»	Will your next degree course result in the same leaving qualification as the degree course we talked about before, or is it another leaving qualification, e.g. Master instead of Bachelor or elementary school teaching qualification instead of Gymnasium teaching qualification?
-99	«	Filtered
-98	«	Don't know
-97	«	Refused
-92	«	Question erroneously not asked
-54		Missing by design
-29	«	Value from the last sub-episode
	»	Value from last-mentioned sub-episode
1		Same leaving qualification
2		Other qualification

	« tg24205_v1 | spVocTrain | tg24205 »	
Label		Point of time decision for master
Text		When did you make the decision for your master degree program?
-54		Missing by design
1		before starting the previous higher education program
2		During the previous higher education program
3		after ending the previous higher education program

	« ts15219_v1 | spVocTrain | ts15219 »	
Label		Vocational qualification
Text	«	Which civil service examination did you take?
	»	Which civil service examinations did you do?
-99	«	Filtered
-98		Don't know
-92	«	Question erroneously not asked
-55		Not determinable

(...)

-54		Missing by design
-20	«	no qualification
	»	Without any qualification
1	«	Completed apprenticeship (commercial, corporate, trade-oriented, agricultural) journey person's or assistant's certificate (skilled worker's certificate), dual vocational education and training
	»	Completion of an apprenticeship (commercial, corporate, trade-oriented, agricultural), journeyman's or assistant's certificate (skilled worker's certificate), dual training
2		Leaving certificate from a school for health care professionals
3	«	Leaving certificate of Berufsfachschule, leaving certificate of a commercial school
	»	Leaving certificate of a Berufsfachschule, leaving certificate of a Handelsschule
4	«	Other type of leaving certificate of the Fachschule
	»	other type of leaving certificate from a Fachschule
5	«	Master's / foreman's certificate
6	«	Technician's certificate
	»	Technician's training certificate
7		Diplom
8	«	Bachelor
	»	Bachelor's degree
9	«	Master
	»	Master's degree
10	«	Diplom from a university of applied sciences (Dipl(FH))
	»	Diplom from a Fachhochschule (Dipl(FH))
11	«	Diplom from a university
	»	University Diplom
12		Bachelor's degree teaching profession
13	«	Bachelor (not for teaching post)
	»	Bachelor's degree (without teaching profession)
14	«	Master teaching post
	»	Master's degree teaching profession
15	«	Master (not for teaching post)
	»	Master's degree (without teaching profession)
16		Magister
17	«	First state examination for teaching post
	»	First state examination teaching profession
18	«	First state examination (not for teaching post)
	»	First state examination (without teaching)
19	«	Second state examination

(...)

	»	Second/Third state examination
20		Doctorate
21		Habilitation
22		Medical specialist
23		Civil service examination for the subclerical class
24		Civil service examination for the clerical class
25		Civil service examination for the executive class
26		Civil service examination for the administrative class
27		IHK (Chamber of Industry and Commerce) examination
28	«	Other leaving qualification
	»	other qualification
29		Other degree from a higher education institute (e.g., ecclesiastical examination, artistic examination)

« **ts15221_v1** | **spVocTrain** | **ts15221** »

Label	«	Aspired vocational education qualification (reconstructed)
	»	aspired vocational training qualification
Text	«	Which civil service examination [final exam for the different classes of German civil service careers] do you/did you want to do?
	»	Which civil service examinations do/did you want to do?
-98		Don't know
-97	«	Refused
-92		Question erroneously not asked
-55		Not determinable
-54		Missing by design
-20	«	no qualification
	»	No degree
1	«	Completed apprenticeship (commercial, corporate, trade-oriented, agricultural) journey person's or assistant's certificate (skilled worker's certificate), dual vocational education and training
	»	Completion of an apprenticeship (commercial, corporate, trade-oriented, agricultural), journeyman's or assistant's certificate (skilled worker's certificate), dual training
2		Leaving certificate from a school for health care professionals
3	«	Leaving certificate of Berufsfachschule, leaving certificate of a commercial school
	»	Leaving certificate of a Berufsfachschule, leaving certificate of a Handelsschule
4	«	Other type of leaving certificate of the Fachschule
	»	other type of leaving certificate from a Fachschule

(...)

5	«	Master's / foreman's certificate
6	«	Technician's certificate
	»	Technician's training certificate
7		Diplom
8	«	Bachelor
	»	Bachelor's degree
9	«	Master
	»	Master's degree
10	«	Diplom from a university of applied sciences (Dipl(FH))
	»	Diplom from a Fachhochschule (Dipl(FH))
11	«	Diplom from a university
	»	University Diplom
12		Bachelor's degree teaching profession
13	«	Bachelor (not for teaching post)
	»	Bachelor's degree (without teaching profession)
14	«	Master teaching post
	»	Master's degree teaching profession
15	«	Master (not for teaching post)
	»	Master's degree (without teaching profession)
16		Magister
17	«	First state examination for teaching post
	»	First state examination teaching profession
18	«	First state examination (not for teaching post)
	»	First state examination (without teaching)
19	«	Second state examination
	»	Second/Third state examination
20		Doctorate
21		Habilitation
22		Medical specialist
23		Civil service examination for the subclerical class
24		Civil service examination for the clerical class
25		Civil service examination for the executive class
26		Civil service examination for the administrative class
27		IHK (Chamber of Industry and Commerce) examination
28	«	Other leaving qualification
	»	other qualification
29		Other degree from a higher education institute (e.g., ecclesiastical examination, artistic examination)

	«	**tg2452m_v1 | spVocTrain | tg2452m** »
Label	«	Start of the doctorate (month)
	»	Starting time of the doctorate (month)
Text	«	And when did you begin the content-related work on your doctorate?
	»	And when have you started with the content work for your doctorate?
-99	«	Filtered
-98		Don’t know
-97		Refused
-96	«	Not in list
-95	«	Implausible value
-94	«	Not reached
-93		Does not apply
-92	«	Question erroneously not asked
-91	«	Survey aborted
-90	«	Unspecific missing
-56	«	Not participated
-55	«	Not determinable
-54		Missing by design
-53	«	Anonymized
-52	«	Implausible value removed
-51	«	No estimate in check module
1	»	January
2	»	February
3	»	March
4	»	April
5	»	May
6	»	June
7	»	July
8	»	August
9	»	September
10	»	October
11	»	November
12	»	December
21	»	Beginning of the year/winter
24	»	Spring/Easter
27	»	Mid-Year/Summer
30	»	Fall
32	»	End of year

	«	tg2452y_v1 | spVocTrain | tg2452y »
Label	«	Start of the doctorate (year)
	»	Starting time of the doctorate (year)
Text	«	And when did you begin the content-related work on your doctorate?
	»	And when have you started with the content work for your doctorate?
-99		Filtered
-98		Don't know
-97		Refused
-96		Not in list
-95		Implausible value
-94		Not reached
-93		Does not apply
-92		Question erroneously not asked
-91		Survey aborted
-90		Unspecific missing
-56		Not participated
-55		Not determinable
-54		Missing by design
-53		Anonymized
-52		Implausible value removed
-51		No estimate in check module