NEPS
National Educational Panel Study

Research Data

FDZ-LIfBi

Data Manual

NEPS Starting Cohort 2—Kindergarten
From Kindergarten to Elementary School

Scientific Use File Version 11.0.0

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Copyrighted Material
Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3, 96047 Bamberg
Director: Prof. Dr. Cordula Artelt
Director of Administration: Dr. Stefan Echinger
Bamberg; July 9, 2024

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. (2024). *Data Manual NEPS Starting Cohort 2–Kindergarten, From Kindergarten to Elementary School, Scientific Use File Version 11.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

This data manual for Starting Cohort 2–Kindergarten "From Kindergarten to Elementary School" has been prepared by the staff of the Research Data Center at the Leibniz Institute for Educational Trajectories (FDZ-LIfBi). It represents a major collaborative effort.

The contribution of the following persons is gratefully acknowledged:

Dietmar Angerer
Daniel Fuß
Tobias Koberg
Gregor Lampel
Sven Pelz
Benno Schönberger
Katja Vogel

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

Daniel Bela, Simon Dickopf, 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/FDZ`

Contents

1 Introduction

1.1 About this manual

This manual facilitates your work with data of the NEPS Starting Cohort 2–Kindergarten (NEPS SC2). It serves both as a first guide for getting started with the complex data and as a reference book. The primary emphasis is on aspects such as sample development, conventions of data preparation, data structure, and merging of information. The manual is neither complete nor exhaustive, but several links to other resources are provided in the respective paragraphs. 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 – this manual is regularly updated and revised for ongoing NEPS starting cohorts.

The first chapter refers to further documentation material, requirements for data access, instructions for data citation, some general rules and recommendations, and selected services provided by the FDZ-LIfBi for NEPS data users. In the second chapter, the fundamental objectives of Starting Cohort 2 and its sampling strategy are briefly introduced. The main part of this chapter describes the sample development across the waves including field times, realized case numbers, survey modes, and the measurement of competence domains. The general principles of Scientific Use File data-editing processes as well as the applied conventions for naming the data files and variables are introduced 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 the relevant data types, identifiers, and short portraits of all available datasets in the Scientific Use File. These portraits also include syntax examples for merging variables of this dataset with variables from other datasets.

The contents of the first chapter as well as large parts of the third and fourth chapters apply to the Scientific Use Files of all NEPS starting cohorts. It is not mandatory that the examples mentioned there explicitly refer to Starting Cohort 2, but they are transferable accordingly.

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 Kindergarten` > `Documentation`

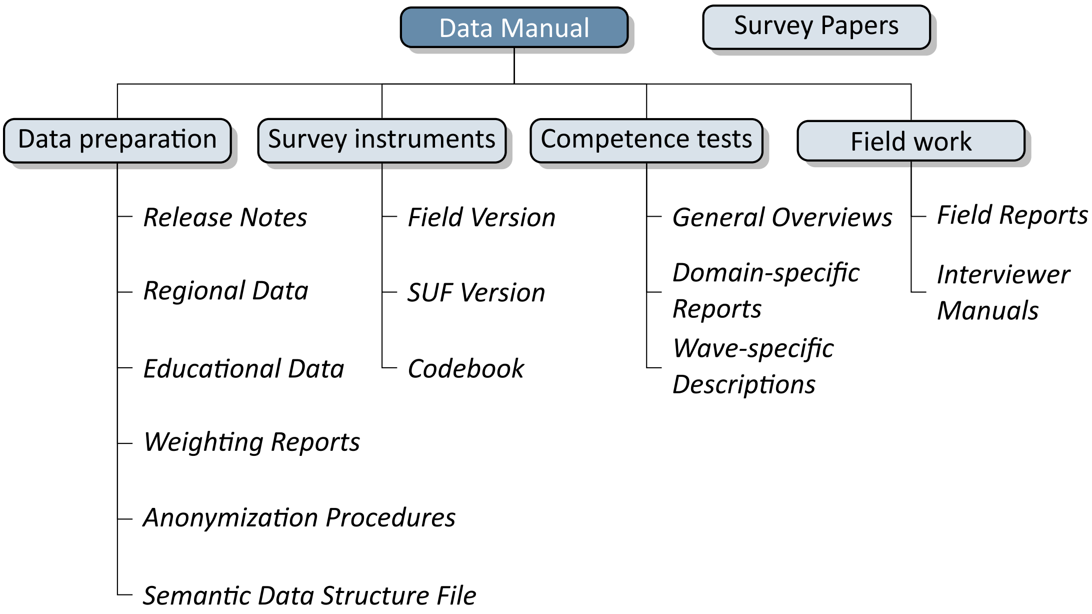

Figure 1: NEPS supplementary data documentation

Release Notes All Scientific Use Files are accompanied by release notes that log changes in the data compared to prior Scientific Use File versions and list bugs eliminated or at least known of. 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.

Educational Data The report gives an overview of the generation of the derived educational variables ISCED, CASMIN and Years of Education.

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 field versions and Scientific Use File (SUF) 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. Usually, for each domain there is a brief description of the construct with sample items as well as 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 and 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 from the LIfBi website at:

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

Additional documentation material might be available for this Starting Cohort. Please visit the documentation website mentioned at the beginning of this chapter 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. Therefore, it is strongly recommended to use the most current release of a Scientific Use File.

File Format

All Scientific Use Files are provided in Stata and SPSS format with bilingual variable 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]
```

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 used 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:SC2:11.0.0`. Other releases of Scientific Use Files for Starting Cohort 2 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 2

SUF Version	DOI	Date of release
11.0.0 (current)	doi:10.5157/NEPS:SC2:11.0.0	**2024-07-09**
10.0.0	doi:10.5157/NEPS:SC2:10.0.0	2022-08-31
9.0.0	doi:10.5157/NEPS:SC2:9.0.0	2020-12-11
8.0.1	doi:10.5157/NEPS:SC2:8.0.1	2020-01-27
8.0.0	doi:10.5157/NEPS:SC2:8.0.0	2019-06-12
7.0.0	doi:10.5157/NEPS:SC2:7.0.0	2018-08-20
6.0.1	doi:10.5157/NEPS:SC2:6.0.1	2018-03-07
6.0.0	doi:10.5157/NEPS:SC2:6.0.0	2017-12-08
5.1.0	doi:10.5157/NEPS:SC2:5.1.0	2017-04-28
5.0.0	doi:10.5157/NEPS:SC2:5.0.0	2017-02-28
4.0.0	doi:10.5157/NEPS:SC2:4.0.0	2016-06-30
3.0.0	doi:10.5157/NEPS:SC2:3.0.0	2015-10-09
2.0.0	doi:10.5157/NEPS:SC2:2.0.0	2013-10-23
1.0.0	doi:10.5157/NEPS:SC2:1.0.0	2012-09-21

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 access 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 all 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 or mail. 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 more sensitive 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 mode 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 requires a guest stay at the LIfBi in Bamberg. More details about the access modes can be found 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 (*Bundeslandkennung*, see Section 1.7) are only available in the controlled server 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 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 2.

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. (2024). *National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC2:11.0.0

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 the NEPS data are requested to notify the Research Data Center about their articles by sending an e-mail with the bibliographic details to `fdz@lifbi.de`. All known 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. (2024). *Data Manual NEPS Starting Cohort 2–Kindergarten, From Kindergarten to Elementary School, Scientific Use File Version 11.0.0*. 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. (2024). *Starting Cohort 2: Kindergarten (SC2), Wave 11, Questionnaires (SUF Version 11.0.0)*. 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 NEPS 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 such a re-identification. 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 mentioned by name in the agreement). 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 single 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 a 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.

Recommendations

In addition to the aforementioned rules, there are some recommendations for using the 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 research 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 before starting the analyses. The work with complex data requires a precise idea of how the information were collected and processed. All relevant material is available online.
- *Facilitated handling of the data:* Use the tools that are offered. Several user services are provided to support NEPS data analyses – 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) and an online discussion forum (e. g., for asking specific 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 the 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 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 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 guest working stations at the LIfBi 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 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 registered NEPS analysis projects, a bibliography with all known publications based on NEPS data, a reference to several NEPS-related events, and a LIfBi data newsletter. All subsequently introduced services and tools can be reached via this website:

→ www.neps-data.de > NEPS

Online Forum

The so-called *Forum4MICA – Making Information Commonly Available* is an open discussion platform for data 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 directly 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 in the forum. Active participation is encouraged and 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 forum at:

→ `https://forum.lifbi.de`

Variable Search

The *Variable Search* facilitates an interactive and quick full text search through all instruments of released NEPS surveys, including competence variables. 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 topic search. The *Variable Search* offers some helpful functions such as displaying occurence in the NEPS starting cohorts, the answering scheme, relevant references, etc. As a web application the service 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` > `Variable Search`

NEPStools

NEPStools is a free to use collection of Stata commands that is created and supplied by the Research Data Center at LIfBi. 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, 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, for the data dissemination, and for the user support including individual advice. We appreciate any feedback in order to further improve our services. This particularly applies to this manual as the guiding document to facilitate your work with the data of Starting Cohort 2.

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

E-mail: `fdz@lifbi.de`
Web: → `www.neps-data.de` > `Data Center` > `Research Data Center`

2 Sampling and Survey Overview

2.1 From kindergarten to elementary school

Starting Cohort 2 takes a longitudinal perspective on children in preschool institutions and elementary school. The study, which was launched nationwide in winter 2010/11, covers educational processes at kindergarten and primary school age in two important stages – kindergarten, including the transition to elementary school, and elementary school, including the transition to lower secondary level (see Berendes et al., 2019).

Describing and explaining the extensive development of cognitive and non-cognitive competencies in early childhood with regard to family conditions and institutional learning environments are at the center of interest. Since nearly all children in Germany attend kindergarten before entering school, a major aim is to investigate the effects of early nonparental care and education on children's outcomes. One key aspect concerns the role of kindergarten and elementary school as nonfamilial educational settings in compensating for early disparities in skills. Another issue addresses the causes and long-term consequences of such disparities.

The influences of early institutional learning environments on developmental progress and outcomes are considered in close connection with the children's home learning environment. Main questions relate to the relative importance of these environments and the mediation mechanisms within and between them. To answer these questions, both structural and quality aspects of the various learning environments are surveyed. An exemplary research topic refers to the role of different care arrangements with respect to the reconciliation between family life and parental participation in the labor market on the children's acquisition of competencies.

Another core topic deals with educational decisions. The first educational decisions are mainly made by the parents, in consideration of the given structural conditions. Of particular importance for the children's educational success is the second transition in the German education system from elementary to secondary school – usually after completing the first four years of schooling. The secondary school system intends for an explicit between-school tracking which forces parents to make very early decisions about their child's future educational trajectory. In addition to the analysis of determinants for these decisions such as the socio-economic status and the subjective evaluation of decision-relevant aspects by the parents or the recommendations by the elementary class teachers, the extent to which early differentiation contributes to the comparatively high level of educational inequality in Germany in terms of competencies, attended tracks, and attained certificates is a crucial research question.

Further dimensions that are taken into account in line with the general approach of the National Educational Panel Study are the role of migration experiences and migration background, the benefits of education in terms of health and well-being, and the personality of the children along with indicators of motivation, self-concepts, interests, and social behavior.

The focus of Starting Cohort 2 is on educational participation and educational processes in preschool age and in the first years of school including the transitions from kindergarten to elementary school and from elementary school to the tracked system of secondary school. To provide appropriate data, a sample was drawn starting with children attending kindergarten two years before they started school. This sample was surveyed year by year and expanded at school entry. In addition to direct competence tests and questionnaires for the children, information was also collected from parents, from kindergarten educators and class teachers in elementary school as well as from the principals.

2.2 Sampling strategy

The target population for the first wave of Starting Cohort 2 consisted of children who attended kindergarten in Germany in 2010/2011 and were expected to start school in the 2012/2013 school year. These children were about four to five years old at the beginning of the panel study and thus usually two years before school enrollment. A special feature of this cohort is that the children are followed through their transition from kindergarten to elementary school. The first two waves of the survey took place in kindergartens, from the third to the sixth wave the survey was continued in elementary schools.

This design posed particular challenges for the sampling strategy. On the one hand, as many of the participating kindergarten children as possible should be “re-found” after their transition to elementary school and further surveyed there. On the other hand, neither for kindergarten children nor for kindergarten institutions was a complete list available to directly draw a sample of target persons. The applied method was **indirect sampling** using the structural link between kindergartens that “transfer” children to elementary schools. This link via institutions was used to get access to the population of kindergarten children. First, a sample of elementary schools was drawn (so-called “NEPS schools”), the selected schools then compiled lists of kindergartens from which they obtain first graders, and from these lists a sample of kindergartens was drawn to finally recruit the sample of target children. The idea and hope were that two years after panel start, a significant proportion of children from the NEPS kindergartens would enter the selected NEPS schools who had already consented to participate in the study during the sampling process. In detail, the two-stage indirect sampling approach was implemented as follows:

In the **first step**, elementary schools were drawn at random and asked to participate in the NEPS survey in the 2012/2013 school year (3rd wave). The basis for the selection was a current and complete list of all 16,824 schools in the general education system (excluding special schools) in Germany as of 2008/09 with pupils in at least one class in the first year. The elementary schools were drawn using a systemic probability proportional to size sampling (pps) and based on implicit stratification according to Federal States, regional classification, and organizing institution/founding. For each of the 400 original schools drawn, four replacement schools with the same characteristics were additionally drawn in order to adequately compensate for dropouts due to refused participation. The challenging school recruitment process began in April 2010 and was completed with 212 elementary schools willing to take part in the study.

In a **second step**, kindergartens were drawn by first identifying the “supplier kindergartens” of the 212 participating elementary schools and then making a selection from these. To identify the eligible kindergartens, the elementary schools were asked to list all kindergartens and the number of children who came to the school from these kindergartens in the 2009/2010 school year. From this list of 1,432 kindergartens (with 9,610 children), a certain number of kindergartens were chosen for each school depending on the size of the list and proportional to the number of listed children. In the end, a total of 981 kindergartens (original and replacement) were available for the recruitment process. This process started in August 2010 with letters sent to the kindergartens and their owners or organizations. 286 kindergartens were recruited to participate in the study, of which 279 finally took part. This corresponds to an average of 1.3 kindergartens per elementary school willing to participate, with a minimum of 0 and a maximum of 5 “supplier kindergartens”.

At the kindergarten level, no further selection of groups took place. Within the sampled institutions, the parents of all children who were of school age for the 2012/2013 school year were asked for their child’s participation. Participation was generally voluntary. The coordinator of each kindergarten handed out the study letters with an attached consent form to the corresponding parents and subsequently collected the completed forms. Children were only allowed to participate if a written consent for the participation of both the child and a parent in the study was available. Thus, panel consent for kindergarten children and parents was coupled. Valid parental consent forms were obtained for 3,007 of the 5,346 children in the target population from the 279 participating kindergartens (56.2%). Due to withdrawals, the size of the initial sample of Starting Cohort 2 was reduced to 2,996 target children, of whom 2,948 took part in the surveys of the first wave (see Figure 2).

Augmentation sample and individual retracking

In 2012, the surveyed kindergarten children moved to elementary school. Children who transitioned to the previously sampled “NEPS schools” were followed up within this institutional context. To augment the sample, all classmates of these children in the first grade were incorporated into the study. An additional augmentation of the sample was achieved by drawing and contacting additional elementary schools, without children from the kindergarten survey being enrolled there.[1] And finally, there are the kindergarten children who changed to a different school (not a “NEPS school”) in 2012. These children were retracked individually, i. e., they skipped waves 3 to 5 as temporary dropouts by design and returned to the tests and surveys in wave 6, which were then conducted in the home context. Thus, three groups of target children can be distinguished in the panel development of Starting Cohort 2.

[1] On top of the 212 participating elementary schools from the initial sampling, another 237 elementary schools were asked to participate in the NEPS study. These schools were selected in a tranched procedure based on the same school list from 2008/09 and on the same sampling principles. A total of 374 schools agreed, constituting a gross sample of 19,205 students in grade 1 in the 2012/2013 school year. All children in this gross sample were invited to join the survey. Accordingly, all parents whose child attended the first grade of one of the participating elementary schools were asked for their and their child’s cooperation. For the third wave, there were 6,917 completed forms with parental consent submitted.

- **Group 1**: The group of kindergarten children being tested in kindergarten in waves 1 and 2 and transitioned to elementary schools sampled in advance and participating. In waves 3 to 6, they are surveyed together with the children of the second group.
- **Group 2**: This group forms the augmentation sample of wave 3. It consists of children that were surveyed and tested in the grades 1 to 4 in elementary schools but were not surveyed or tested in kindergarten institutions in waves 1 and 2.
- **Group 3**: This group includes the kindergarten children who could not be ”re-found” in the previously selected elementary school that were willing to participate. Contact to these children was maintained through the annual survey of parents; from wave 6 onwards, the children of this group were interviewed and tested individually at home.

Groups 1 and 3 constitute the kindergarten panel. Groups 1 and 2 jointly represent the first-grade panel that is followed within the institutional context. As Figure 2 symbolizes, only a small proportion of the target children could be followed through the institutional transition (Group 1). All three groups were surveyed and tested together again in 2015, in wave 6 at the time of the fourth grade – regardless of whether they were in the institutional context of a “NEPS school” or not. For the children in Group 1 and 2, this was the last survey in the school context; for the children in Group 3, this survey took place in a similar time corridor at home. From wave 7 onwards, all students were tested and interviewed individually together with their parents. The groups can be distinguished by the variable `tx80115` (“Sample Group”) in `CohortProfile`.

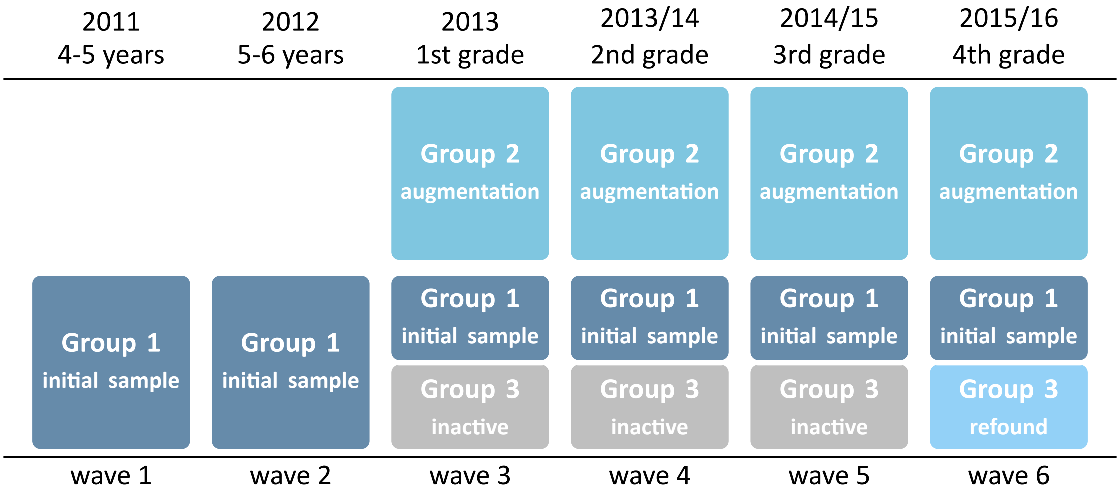

Figure 2: Panel development of children of Starting Cohort 2

The sampling design and its consequences for the derivation of sampling weights are fully described in Steinhauer and Zinn, 2016, Zinn et al., 2018, and Aßmann et al., 2019. Further remarks on the recruiting process are given in the PAPI field reports of the respective survey waves (in German only, see Section 1.2).

Context persons

Target persons of Starting Cohort 2 are children, beginning with the first survey in kindergartens or in the case of the augmentation or refreshment sample two years later in grade 1 at regular schools. In order to collect additional information about the children and the **institutional learning environments**, contextual data were collected from the kindergarten educators, the class teachers, as well as from the kindergarten and the school principals. These supplementary surveys were conducted using self-administered paper questionnaires (PAPI). Participation was also voluntary for all context persons. In survey waves 1 to 6, *educators and teachers* filled a short assessment questionnaire for each participating child and a general questionnaire about the class context, teaching activities, own educational trajectory, and further personal topics. At the same time, the *institution management* each answered a questionnaire for kindergarten- or school-related information and a few questions about themselves. Please note that the dataset with the last-mentioned information from the institution management (`xInstitution`) is only available in the RemoteNEPS and On-Site version of the Scientific Use Files.

The **family or home learning environment** of the target children was covered by repeated interviews with a parent or legal guardian. These interviews took place in waves 1 to 7 as well as in wave 9 and wave 11. It has to be emphasized that the parental interviews also took place in waves 3 to 5 for those target children who did not transition to a “NEPS school” and therefore did not participate in the survey until wave 7. The collection of information from the parents was conducted in the form of computer-assisted telephone interviews (CATI). Whenever possible, the biological or social parent who knew most about the child’s school affairs was interviewed. In more than three quarters of the cases, this was the biological mother. During the course of the panel, it was possible to switch to another parent with parental authority. Participation in the interviews was voluntary. The *parent* interviews focused primarily on educational aspects and the school history of the children, but also on parental support for the children, children’s health, satisfaction with school, language use in the family, household equipment and various socio-demographic characteristics.

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

2.3 Competence measures

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

Data from the competence tests pass through an editing process before they get integrated into the Scientific Use File. This data preparation enables users to work with scored items and generated test scores such as the sum or mean of correct answers. Detailed descriptions on how these scores were estimated can be found in separate reports for the respective competence domains (see Section 1.2). The individual and generated scores are compiled in the dataset named `xTargetCompetencies`.[2] 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 (see Section 4). As a consequence, variable names for competence scores follow a specific nomenclature. These conventions not only allow for the identification of the respective domain, the target group, the testing modus, and the kind of scoring, they also inform about the repeated administration of a test item in a different wave or starting cohort (see Section 3.2.2).

The next table shows the schedule of competence measures in Starting Cohort 2 with domains by waves and test modes.

- Subsequent to several competence tests (rx/re, vo, gr, ma, sc, nr/nt, ic, or), the target children had to assess their own test performance (*Procedural Metacognition*, mp).[3]
- The L1-Test for *Russian and Turkish Language* (nr/nt) has been applied to target children of a corresponding migration background only.
- Reduced testing:
  - For individually traced target children, competence tests were realized in the domains of *Reading* (re) and *Mathematics* (ma) only in wave 6. Tasks relating to the domains *Orthography* (or) and *Delayed Gratification* (de) were not administered to this group of children.
  - In wave 9, a randomized allocation of competence tests with two out of the three domains (re + ma OR re + sc OR ma + sc) has been applied.
  - Only the DGCF subtest *Reasoning* was administered in wave 11, but not *Perceptual Speed*.
- In the survey waves 7, 8, and 10, there were no competence tests at all administered to the target children.

[2] The Scientific Use File contains another competence dataset called `xPlausibleValues`, which contains exemplary variables with plausible values that were generated using the freely available R package *NEPSscaling* (see Scharl and Zink, 2022 and Section 1.8).

[3] The list of all NEPS competence domains together with the respective abbreviation can be found in Table 5.

Table 2: Schedule of competence measures. P = Paper-Based Test (proctored)

		2011	2012	2013	2013/14	2014/15	2015/16	2018/19	2021
		Wave 1	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**	**Wave 9**	**Wave 11**
		4-5 years	5-6 years	Grade 1	Grade 2	Grade 3	Grade 4	Grade 7	Grade 9
Domain-General Competencies									
DGCF: Cognitive Basic Skills	`dg`	—	P	—	P	—	—	—	P
Domain-Specific Competencies									
(Early) Reading Competence	`rx/re`	—	—	—	P	—	P	P	—
Reading Speed	`rs`	—	—	—	P	—	—	—	—
Vocabulary: LC at Word Level	`vo`	P	—	P	—	P	—	—	—
Grammar: LC at Sentence Level	`gr`	P	—	P	—	—	—	—	—
Mathematical Competence	`ma`	—	P	P	P	—	P	P	P
Scientific Competence	`sc`	P	—	P	—	P	—	P	—
Native Language Russian/Turkish: LC	`nr/nt`	—	—	—	P	—	—	—	—
Metacompetencies									
Declarative Metacognition	`md`	—	—	P	—	P	—	—	—
ICT Literacy	`ic`	—	—	—	—	P	—	—	—
Stage-Specific Competencies									
Early Knowledge of Letters	`lk`	—	P	—	—	—	—	—	—
Phonological Working Memory	`ds/bd`	—	P	—	—	—	—	—	—
Phonological Awareness	`on/ri/ip`	—	P	—	—	—	—	—	—
Delayed Gratification: Executive Control	`de`	—	P	—	—	—	P	—	—
Orthography	`or`	—	—	—	—	—	P	—	—

2.4 Survey overview and sample development

This section informs about the progress of the Starting Cohort 2 sample. For each survey wave in the current Scientific Use File, there is a short characterization in terms of field time, groups of respondents, number of realized cases, survey modes, and the survey institute(s) responsible for collecting the data. A more detailed insight into all aspects of the field work can be found in the wave-specific *Field Reports*, which are available on the website (in German only) as part of the data documentation.

→ `www.neps-data.de` > `Data Center` > `Data and Documentation` > `Starting Cohort Kindergarten` > `Documentation`

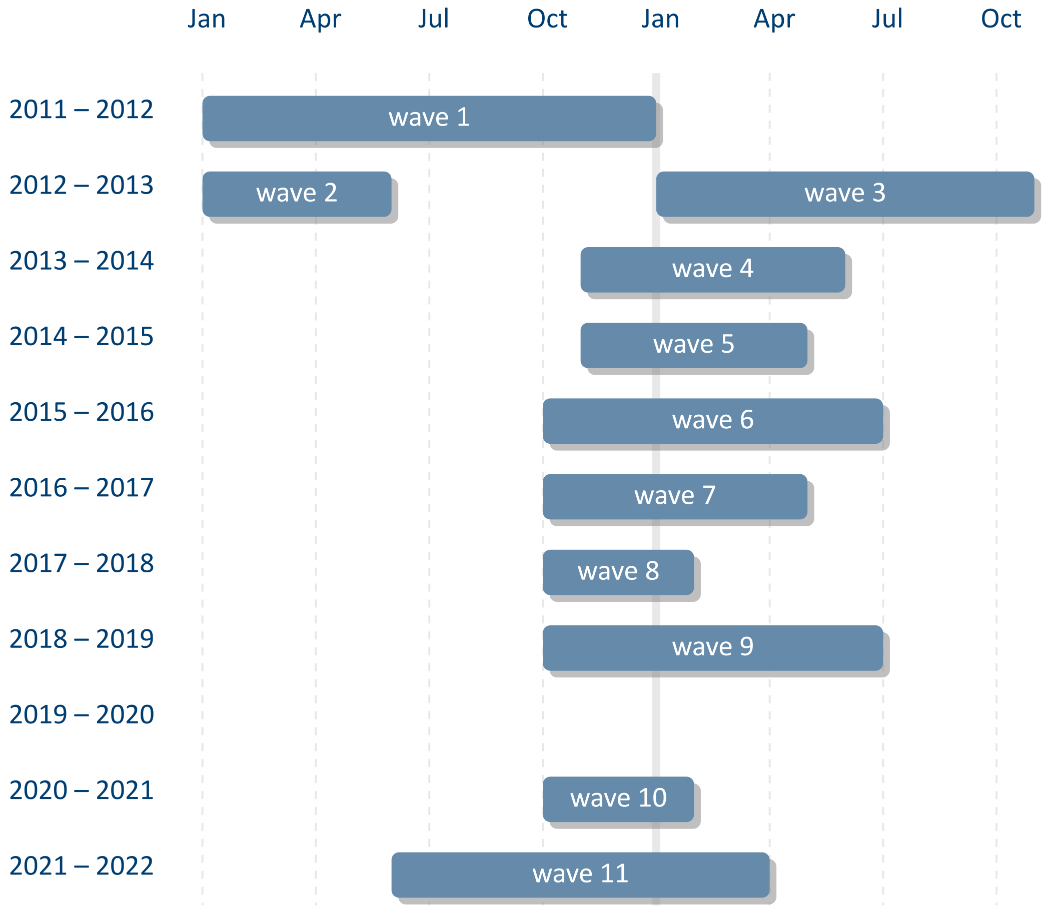

Figure 3: Field times of the survey waves of Starting Cohort 2

2.4.1 Wave 1: 2011

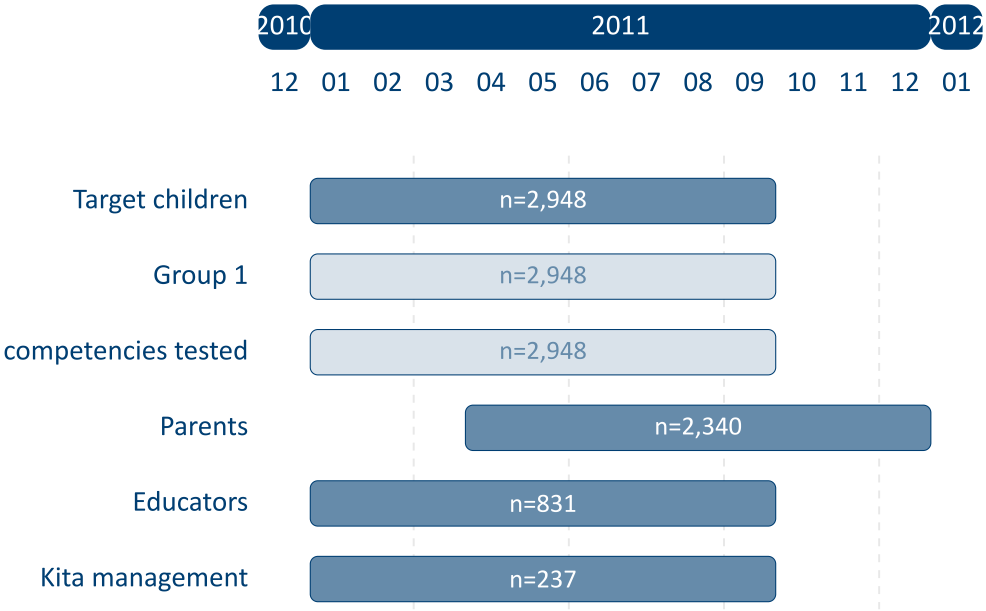

Figure 4: Field times and realized case numbers in wave 1

- **Target persons**
  - *Kindergarten children two years before starting school (approx. 4-5 years old)*

    Inital Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Sampling Indirect sampling of kindergartens via sampling of regular schools at elementary level (see Section 2.2 for details):

    1. Random selection of elementary school from total list for Germany according to size-proportional sampling approach (pps) with implicit stratification by federal states, regional classification and organizing institution/sponsorship
    2. Random selection of kindergartens from a list of “supplier kindergartens” compiled by the participating schools from the first sampling stage; selection size-proportional to the number of children who have transitioned from the kindergarten to the school in the past
    3. All children in the selected kindergartens who were of school age for the 2012/13 school year are included in the initial sample.

Modus Two individual assessments per child in the kindergarten to conduct the competence tests in an age-appropriate, playful manner; the results were recorded by the respective test supervisor using protocol sheets, which also allowed to record non-verbal answers (paper-based)

Competencies Science, Vocabulary, Grammar

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)
  - *Educators*

    Sample Educators responsible for the participating children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the kindergarten group and about oneself (PAPI)
  - *Kindergarten management*

    Sample Heads of the participating kindergartens

    Modus Written questionnaires for contextual information about the kindergarten and about oneself (PAPI)
- **Data collection**
  - *Commercial survey institutes*

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

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

2.4.2 Wave 2: 2012

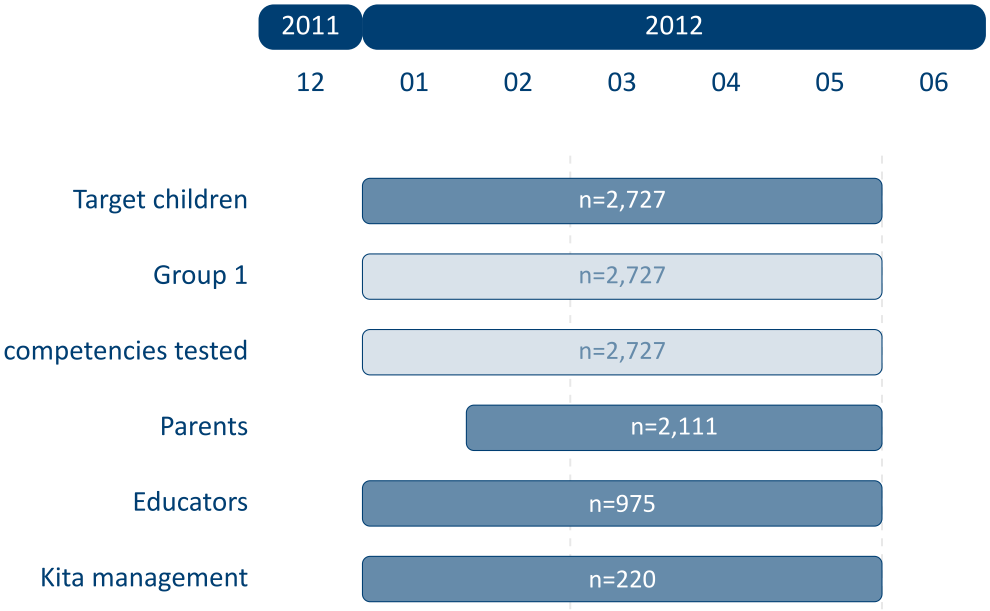

Figure 5: Field times and realized case numbers in wave 2

- **Target persons**
  - *Kindergarten children one year before starting school (approx. 5-6 years old)*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Modus Two individual assessments per child in the kindergarten to conduct the competence tests in an age-appropriate, playful manner; the results were recorded by the respective test supervisor using protocol sheets, which also allowed to record non-verbal answers (paper-based)

    Competencies Cognitive Basic Skills (DGCF), Mathematics and the Stage-specific competencies of Delayed Gratification, Early Knowledge of Letters, Phonological Working Memory, Phonological Awareness

- **Context persons**
  - *Parents*

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

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

  - *Educators*

    Sample Educators responsible for the participating children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the kindergarten group and about oneself (PAPI)

  - *Kindergarten management*

    Sample Heads of the participating kindergartens

    Modus Written questionnaires for contextual information about the kindergarten (PAPI)

- **Data collection**

  - *Commercial survey institutes*

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

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

2.4.3 Wave 3: 2013

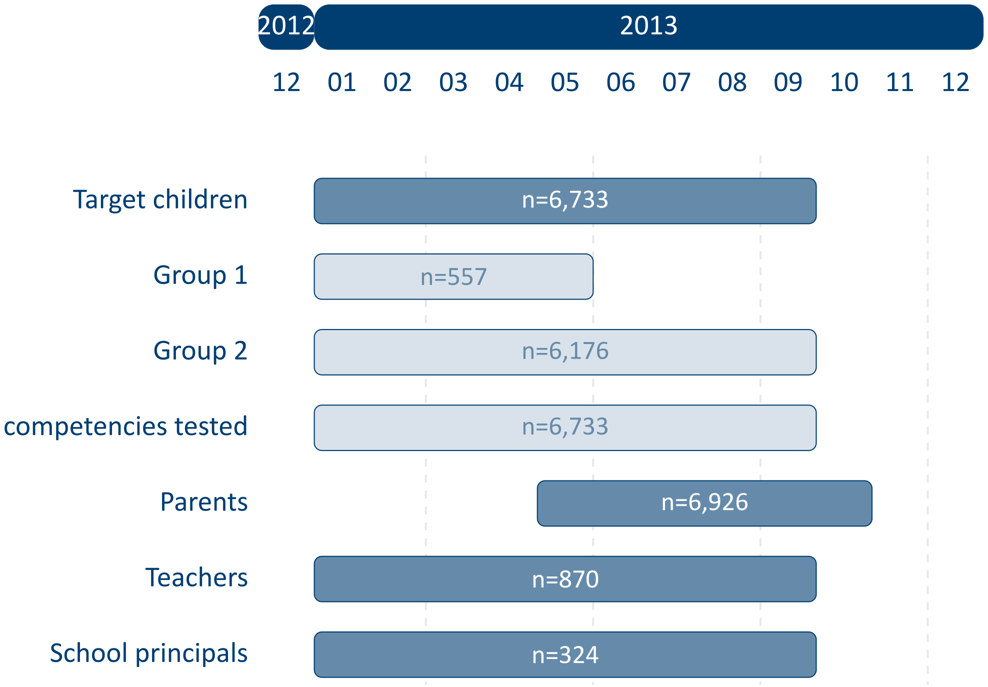

Figure 6: Field times and realized case numbers in wave 3

- **Target persons**
  - *Children in grade 1 at regular schools*

Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave) Children in grade 1 in the 2012/13 school year from additionally sampled schools and from initially sampled schools without having been part of the kindergarten surveys

Sampling Direct sampling of regular schools at elementary level (see also Section 2.2):

1. Random selection of elementary school from the same total list for Germany as for the initial sample and according to the same sampling approach
2. Broadening of the initial sample by including all first graders at the originally selected "NEPS schools" that were willing to participate
3. All children at the selected and participating elementary school who attended first grade in the 2012/13 school year are included in the initial sample.

Modus Written competence tests completed in class context (PAPI)

Competencies Mathematics, Science, Vocabulary, Grammar, Declarative Metacognition

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial sample, regardless of the target child's participation in the school survey (if possible, the same person as in the previous wave, but changing the informant is possible)
    One biological or social parent with parental responsibility per target child from the refreshment sample

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

  - *Teachers*

    Sample Class teachers of the target children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the class and about oneself (PAPI)

  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires for contextual information about the school and about oneself (PAPI)

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

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

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

2.4.4 Wave 4: 2013/2014

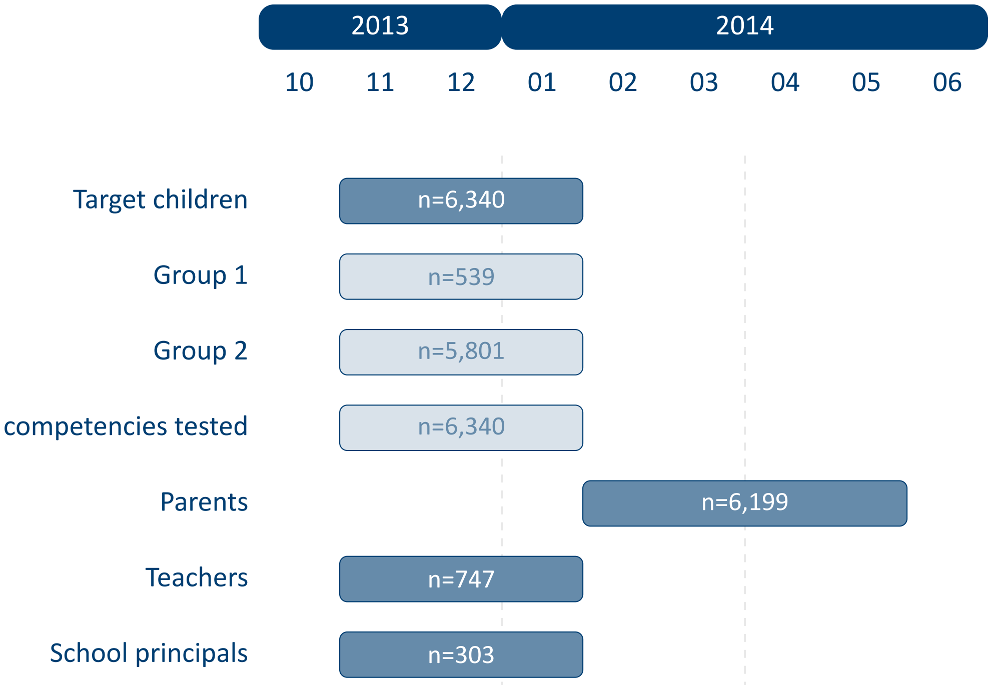

Figure 7: Field times and realized case numbers in wave 4

- **Target persons**
  - *Children in grade 2 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

    Modus Written competence tests completed in class context (PAPI)

    Competencies Cognitive Basic Skills (DGCF), Early Reading, Reading Speed, Mathematics, Native Language Russian or Turkish (only for children with corresponding migration background)

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)
  - *Teachers*

    Sample Class teachers of the target children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the class and about oneself (PAPI)
  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires for contextual information about the school and about oneself (PAPI)
- **Data collection**
  - *Commercial survey institutes*

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

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

2.4.5 Wave 5: 2014/2015

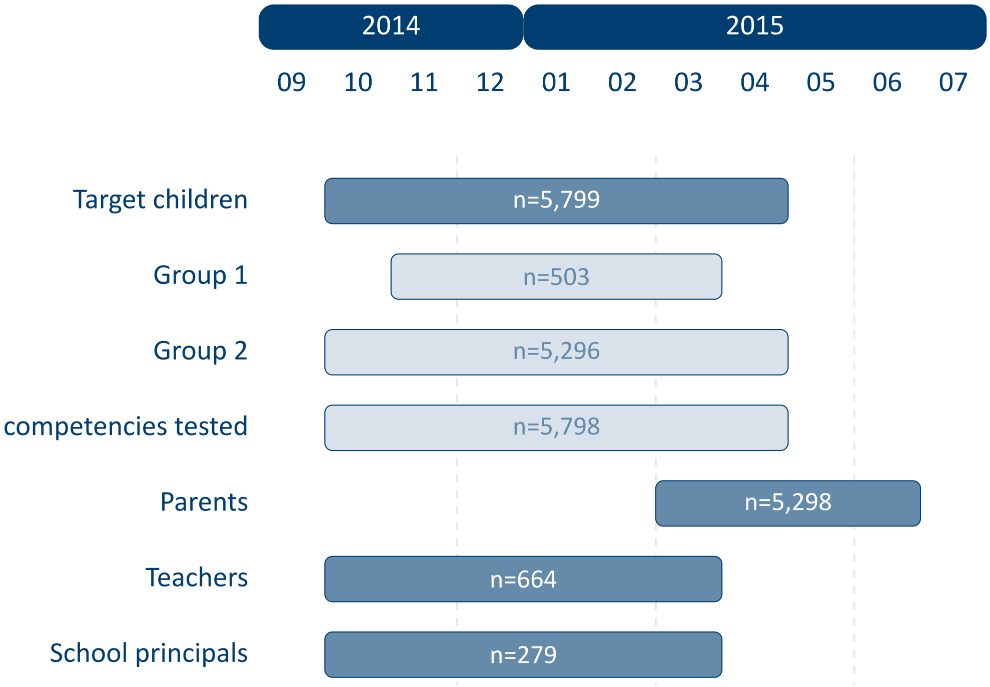

Figure 8: Field times and realized case numbers in wave 5

- **Target persons**
  - *Children in grade 3 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

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

    Competencies Vocabulary, Science, ICT Literacy, Declarative Metacognition

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

  - *Teachers*

    Sample Class teachers of the target children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the class and about oneself (PAPI)

  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires for contextual information about the school and about oneself (PAPI)

- **Data collection**

  - *Commercial survey institutes*

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

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

2.4.6 Wave 6: 2015/2016

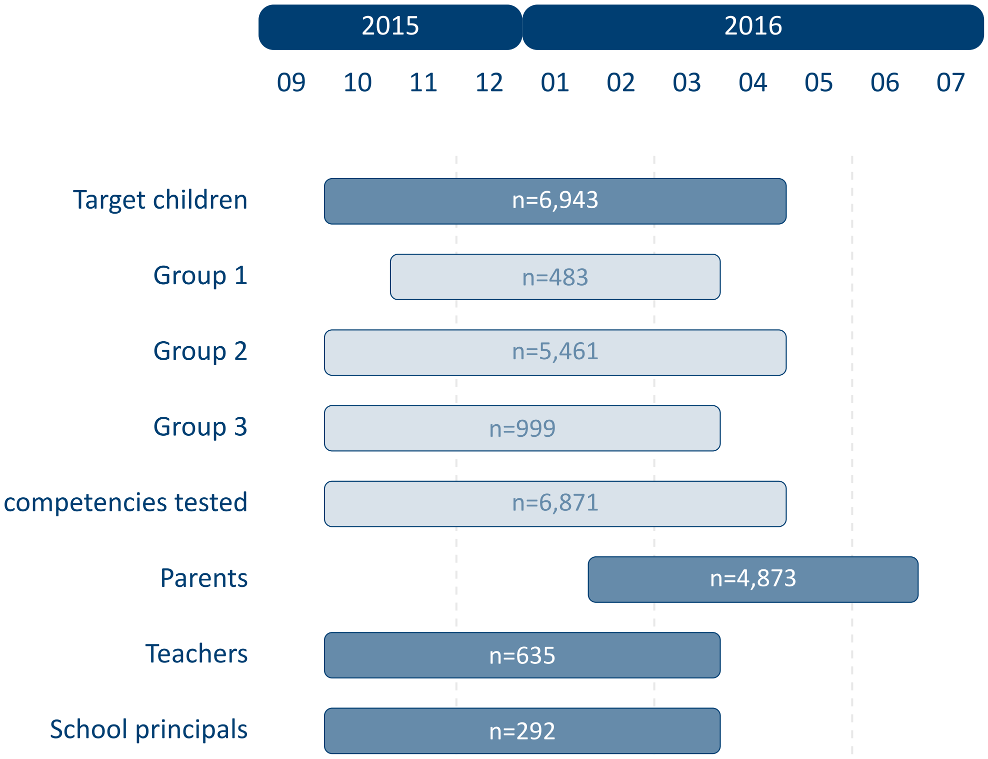

Figure 9: Field times and realized case numbers in wave 6

- **Target persons**
  - *Children in grade 4 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

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

    Competencies Mathematics, Reading and the Stage-specific competencies of Delayed Gratification, Orthography

  - *Individually tracked children in grade 4 at regular schools*

    Subsample Children from the initial or refreshment sample who could not or could no longer be followed in the school context (including the target persons who were not surveyed in schools after the first two waves in kindergartens)

Modus Written questionnaires and competence tests completed in family context with the support of interviewers (PAPI/CAPI)

Competencies Mathematics, Reading

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

  - *Teachers*

    Sample Class teachers of the target children

    Modus Written questionnaire for individual assessment of each target child (PAPI)

    Modus Written questionnaires for information about the class and about oneself (PAPI)

  - *School principals*

    Sample Principals of all schools with participating classes

    Modus Written questionnaires for contextual information about the school and about oneself (PAPI)

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

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

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

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

2.4.7 Wave 7: 2016/2017

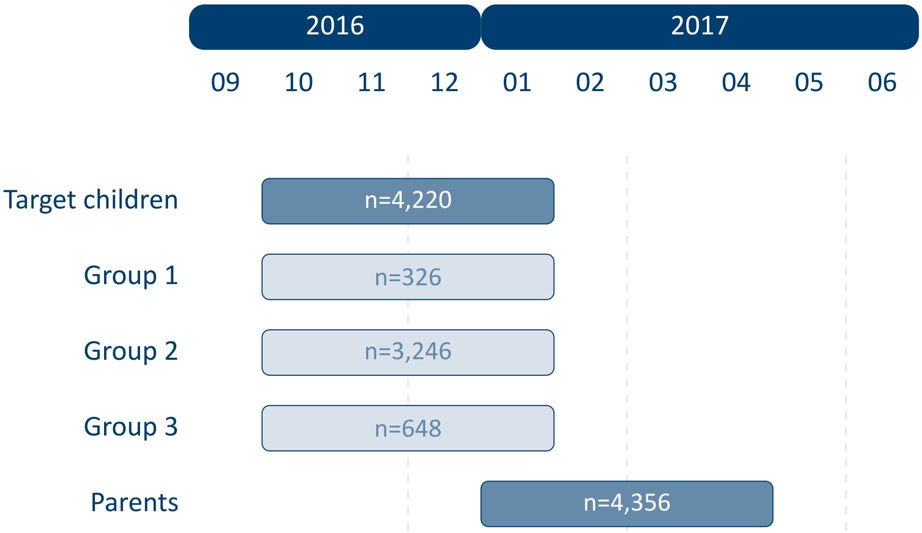

Figure 10: Field times and realized case numbers in wave 7

- **Target persons**
  - *Individually tracked children in grade 5 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

    Modus Written questionnaires completed in family context or online questionnaire as alternative (PAPI/CAWI)

    Competencies No testing in this wave
- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the previous wave, but changing the informant is possible)

    Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

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

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

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

2.4.8 Wave 8: 2017/2018

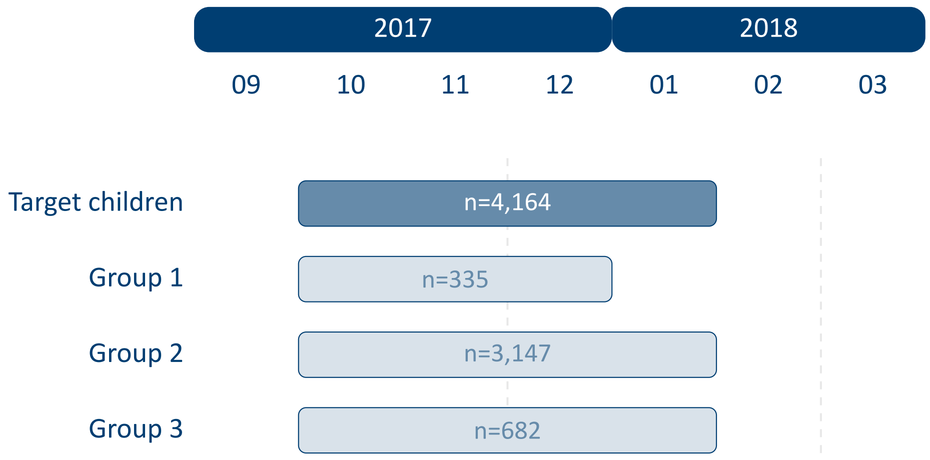

Figure 11: Field times and realized case numbers in wave 8

- **Target persons**
  - *Individually tracked children in grade 6 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

    Modus Written questionnaires completed in family context or online questionnaire as alternative (PAPI/CAWI)

    Competencies No testing in this wave
- **Data collection**
  - *Commercial survey institutes*

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

2.4.9 Wave 9: 2018/2019

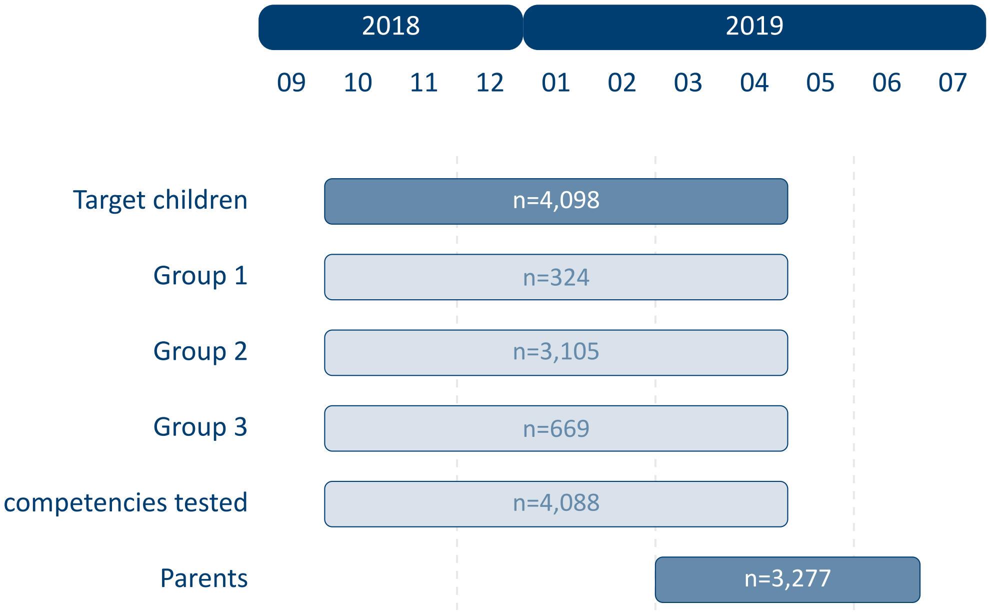

Figure 12: Field times and realized case numbers in wave 9

- **Target persons**
  - *Individually tracked children in grade 7 at regular schools*

Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

Modus Written questionnaires and competence tests completed in family context with the support of interviewers (PAPI/CAPI) or online questionnaire as alternative for questionnaire (CAWI)

Competencies Mathematics, Science, Reading (two of the three domains per child)

- **Context persons**
  - *Parents*

Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the last interview, but changing the informant is possible)

Modus Computer-assisted telephone interviews in German, Russian and Turkish (CATI)

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

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

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

2.4.10 Wave 10: 2020/2021

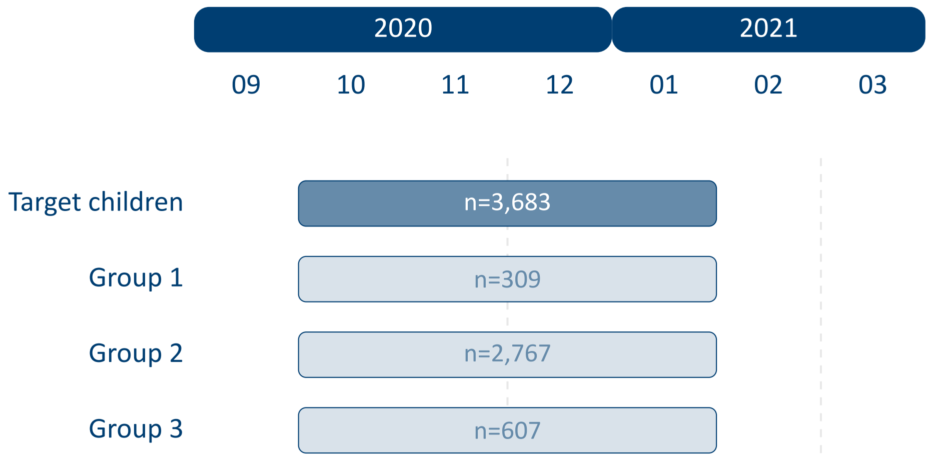

Figure 13: Field times and realized case numbers in wave 10

- **Target persons**
  - *Individually tracked children in grade 9 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

    Modus Written questionnaires completed in family context or online questionnaire as alternative (PAPI/CAWI)

    Competencies No testing in this wave
- **Data collection**
  - *Commercial survey institutes*

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

2.4.11 Wave 11: 2021/2022

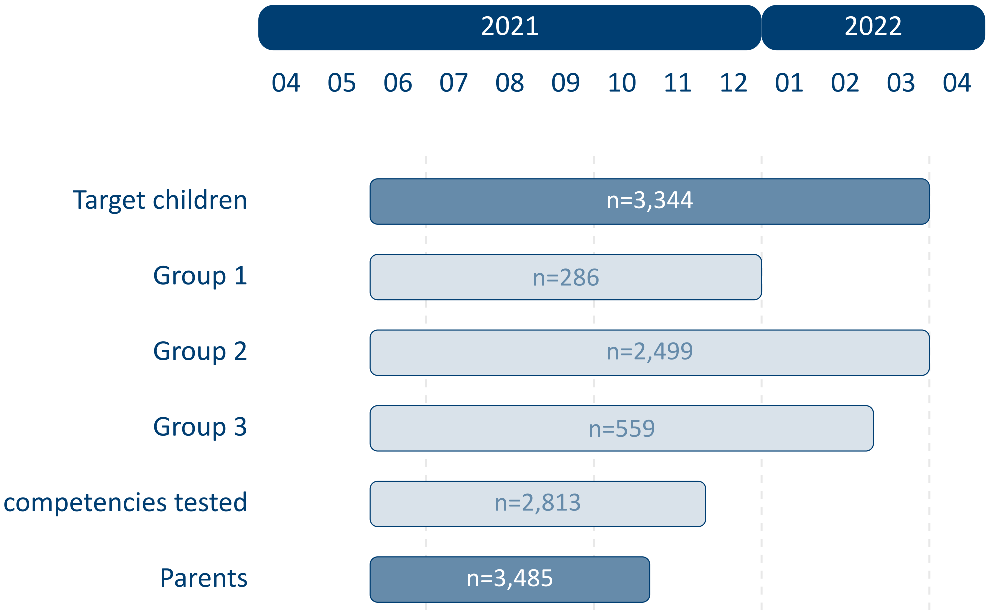

Figure 14: Field times and realized case numbers in wave 11

- **Target persons**
  - *Individually tracked children in grade 9 at regular schools*

    Initial Sample Children who attended kindergarten at panel start in 2011 and reached school age in the 2012/13 school year

    Refreshment Sample Children in grade 1 in the 2012/13 school year (at the time of the third survey wave)

    Modus Written questionnaires and competence tests completed in family context with the support of interviewers (PAPI/CAPI) or online questionnaire as alternative for questionnaire (CAWI)

    Competencies Cognitive Basic Skills (DGCF, only Reasoning), Mathematics

- **Context persons**
  - *Parents*

    Sample One biological or social parent with parental responsibility per target child from the initial and the refreshment sample (if possible, the same person as in the last interview, but changing the informant is possible)

    Modus Computer-assisted telephone interviews in German (CATI)

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

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

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

3 General Conventions

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

1. **The first paradigm is that of unaltered data.** Wherever possible, the content of the original data is neither changed nor modified for the Scientific Use File. This paradigm is the basis for preserving the full research potential of the data collected. Therefore, no corrections are made during data preparation in order to "establish" any content validity. This means that the Scientific Use File may contain implausible values unless appropriate checks were already implemented in the survey instrument. Only in rare cases, in which the responsible developers of a variable request the removal of clearly implausible information in the data, these values are replaced by the special missing code "implausible value removed" (-52, see Table 6). The only systematic exception to this paradigm concerns the recoding of open-ended responses that can be subsequently assigned to a closed response category for the respective question (see Section 3.4 for details). The NEPS Scientific Use Files are provided with a special dataset `EditionBackups` that contains backup information for all content that has been modified by such recoding procedures (see Section 4.5.2 for details).
2. **The second paradigm is to integrate the data as much as possible without compromising the usability of the Scientific Use File.** For this purpose, the original data – some of which comprise over a hundred individual datasets – are combined into a few dozen panel and episode datasets (see Section 4.3 and Section 4.4 for details). This strategy is based on the assumption that it is far more convenient for the vast majority of data users to reduce already integrated data for a specific analysis than to correctly merge the information relevant for the analysis from scattered source data themselves.

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 all NEPS cohorts. Thus, a researcher who is familiar with the data logic of a particular cohort should be able to immediately recognize this structure when starting to work with data from another cohort. The conventions described in the following sections apply equally to Starting Cohort 2, although some of the examples refer to other NEPS cohorts.

3.1 File names

The naming of the data files in the NEPS Scientific Use Files is determined by a few rules that are summarized in Table 3. The four different elements of a dataset name are each separated by an underscore (_).

Table 3: Naming conventions for NEPS data files

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 pTarget contains longitudinal panel data with reference to the target persons; 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

[#]-[#]-[#]	**Indicator for the release version** *First digit*: the main release number is incremented with every further survey wave available; e. g., the first digit 10 implies that data of the first ten waves are included in the Scientific Use File *Second digit*: the major update number is incremented with every bigger change to the Scientific Use File; major updates affect the data structure (updating of analysis syntax 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 or labels (updating of analysis syntax is not necessary)

For instance, the file SC2_CohortProfile_D_11.0.0.dta refers to the generated *CohortProfile* data of *Starting Cohort 2* in its *Download* version of the current Scientific Use File release *11.0.0*.

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 15. 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.

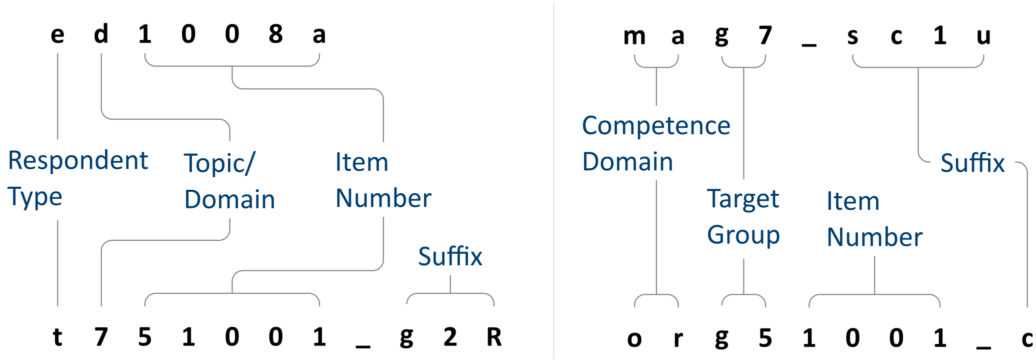

Figure 15: General variable naming (left) and competence variable naming (right)

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 `_v1` for the data before the question was modified for the first time, `_v2` for the data before the question was modified for a second time, and so on.

- *Harmonized variables:* The suffix `_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 _w1, _w2, …, _w10 may mean that there are up to 10 values for this variable per person or episode. This is the case, for example, if the corresponding item in the survey instrument was repeatedly measured in a loop. Another example concerns the date of the competence measurement within a survey wave if it took place on two different days.
- *Confidentiality level:* The _D, _R, or _O suffix indicates variables that have been modified during the anonymization process (see Section 1.4). The suffix _O signalizes that data in this variable is only available via On-site acces; _R refers to variables where access to detailed information is only possible via RemoteNEPS and On-site stay; and _D means that data in this variable has been extracted from the corresponding _O or _R variable to make at least some information available in the Download version of the Scientific Use File. The confidentiality suffixes stand either alone (e. g., country of birth: t405010_R) or in combination with other suffixes (e. g., district of place of birth: t700101_g3R).

3.2.2 Conventions for competence variable naming

The naming of variables from competence measurements and direct measures follows an alternative logic. In contrast to other data files, the competence datasets (xTargetCompetencies and xPlausibleValues, plus xDirectMeasures in Starting Cohort 1) are structured in WIDE format; that is, all values for a single respondent are represented in one row of the data matrix. Thus, the integration of information from several competence domains collected across several survey waves requires specific conventions for variable naming. Competence variables are characterized by three name components and supplementing suffixes. The first component indicates the competence domain of the measurement (two characters, e. g., vo for vocabulary). The second part identifies the target group and the survey wave or class level in which the measurement was first used (two or three characters, e. g., k1 for kindergarten children during the first wave). The target group identification does not necessarily indicate the cohort or testing wave of the measurement. Please refer to the explanations in the next section for the special features of repeatedly used test items. Some competence measurements 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.[4]

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",

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

value 1 = “solved”). In addition to the single 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 in competence variable names 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
`cl`	Civic Literacy
`dc`	Digital competence
`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: Attention abilities
`gk`	General knowledge
`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/discourse level
`lk`	Early knowledge of letters
`ma`	Mathematical competence
`mb`	Mathematical competence (IQB Trends in Student Achievement)
`md/mp`	Declarative metacognition/Procedural metacognition
`ni`	Nonverbal reasoning
`nr/nt`	Native language Russian/Turkish: Listening comprehension
`on`	Blending of onset and rimes: Phonological awareness
`or`	Orthography
`rb`	Reading competence (IQB Trends in Student Achievement)
`re`	Reading competence
`ri`	Rimes: Phonological awareness

(...)

Table 5: (continued)

`rs`	Reading speed
`rx`	Early reading competence
`sc`	Scientific competence
`st`	Scientific thinking: Science propaedeutics
`vi`	Verbal reasoning
`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 the 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: **The meta information available in a dataset always corresponds to the most recent 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., competence test measures)
–95	implausible value	assigned by survey agency (e. g., multiple answers to a one-answer question in PAPI)
–97	refused	as default answer option to the question
–98	don’t know	as default answer option to the question
–20,...,–29	*various*	item-specific missing with informative value label (e. g., “no grade received” for question about school grades)

(...)

Table 6: (continued)

Code	Meaning	Note
Not applicable		
–54	missing by design	question not included in (sub)sample-specific instrument (e. g., not asked in all waves)
–90	unspecific missing	e. g., question not answered, empty field (PAPI)
–91	survey aborted	respondent has quit the interview (CAWI)
–92	question erroneously not asked	question not asked by mistake (CAWI/CATI)
–93	does not apply	as default answer option to the question
–99	filtered	filtered out question (other than CATI/CAPI)
.	*system*	filtered out question (CATI/CAPI)
Edition missings (recoded into missing)		
–52	implausible value removed	only in exceptional cases (at the request of 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-ended 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. Therefore, the recoding 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 scheme 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 16 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 Pelz, 2023.

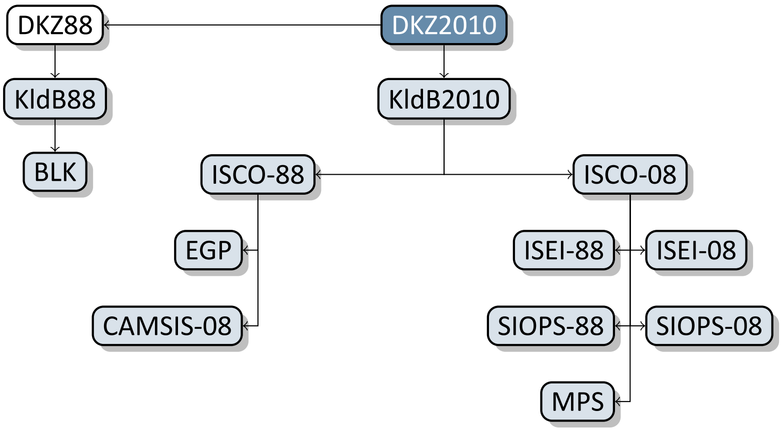

Figure 16: Derivation paths for several occupational scales and schemes provided in the NEPS

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 Files of all NEPS starting cohorts .

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 Scientific Use File. For example, there are competence measurements that were repeatedly carried out with the same target persons. 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. These datasets are marked by a preceding ***sp***. An example is the file `spEmp` in most NEPS starting cohorts, which informs about current and former episodes of employment.

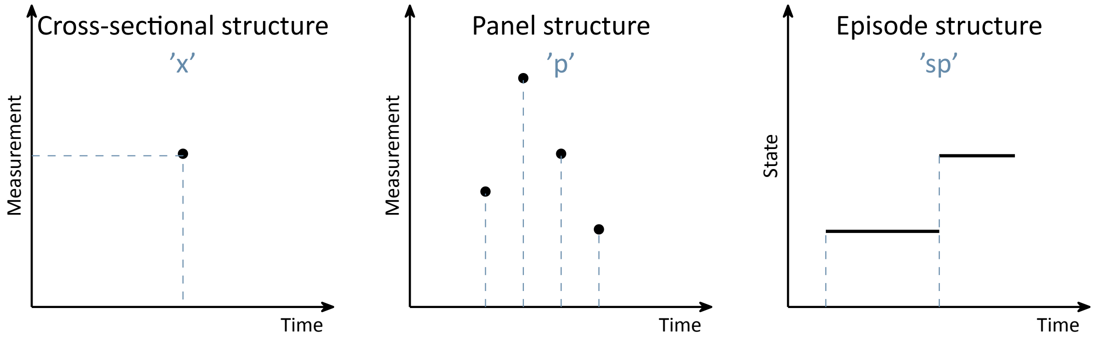

Figure 17: Different types of data structures

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 or CohortProfile, see Figure 19).

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 2 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 biographical data from generated or single episode datasets.

ID_group, ID_cc identifies the kindergarten group and the class in school respectively within a certain wave. These identifier variables are **not** unique across waves.

ID_e uniquely identifies an educator/teacher across waves. This identifier variable can be used to merge data from educators/teachers with observations from children/students. However, it is not possible to merge the data directly with the ID_t (e. g., in the CohortProfile dataset). The linking with data of the target persons or parents or institutions requires the "path" via the group, class or course identifier (ID_group in pGroups or ID_cc in pCourseClass or ID_cg in pCourseGerman or ID_cm in pCourseMath). A concrete example of the procedure is given in Section 4.5.5.

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 the 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

A major focus of the National Educational Panel Study is on recording biographical trajectories as completely as possible. Depending on the NEPS cohort, different areas of the life course are surveyed as so-called ***episodes***. These areas range from school history, education and employment history to household-related histories (e. g., partnership, siblings, children). The retrospective collection of biographical information – What has happened in a certain area of life since time X or since the last interview? When did an episode start and when did it end? What are the characteristics of this episode? – is very demanding and the resulting data material is very complex. Episode or spell data are therefore a particular challenge for the analysis. The following explanations help to better understand this data format and its processing in order to handle it in a meaningful and appropriate way. The information applies equally to all NEPS cohorts, even if the specific data material differs from starting cohort to starting cohort according to the surveyed biographical areas. Information on how to work with the spell data can also be found in the video tutorials offered and in the online forum (see Section 1.2).

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 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". This allows episodes with incomplete dates to be included in the plausibility test during the interview (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 concern the `Biography` dataset only. In the spell datasets for the different life domains, the information provided by respondents during the interview with regard to the start and end dates of episodes remains unchanged.

- Firstly, 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 specified. 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 unchanged.
- Secondly, one- and two-month gaps between consecutive episodes are 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`, close larger gaps completely.

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 respondent in order to recall 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 practically 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.

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 additionally summarized in a separate data line (record) according to defined rules. This data line reflects the most current or relevant information of the entire episode, depending on the harmonization rule applied (see below in this chapter). This ususally 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 time of the last interview, the information valid at the time of the last interview is selected. We call this process of summarizing information about an episode from different survey waves ***episode harmonization***. It is described in detail below.

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

If an episode starts and ends within the retrospectively queried time period of a survey wave (spell 1 in interview A, see Figure 18), 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 been 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 18). 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).

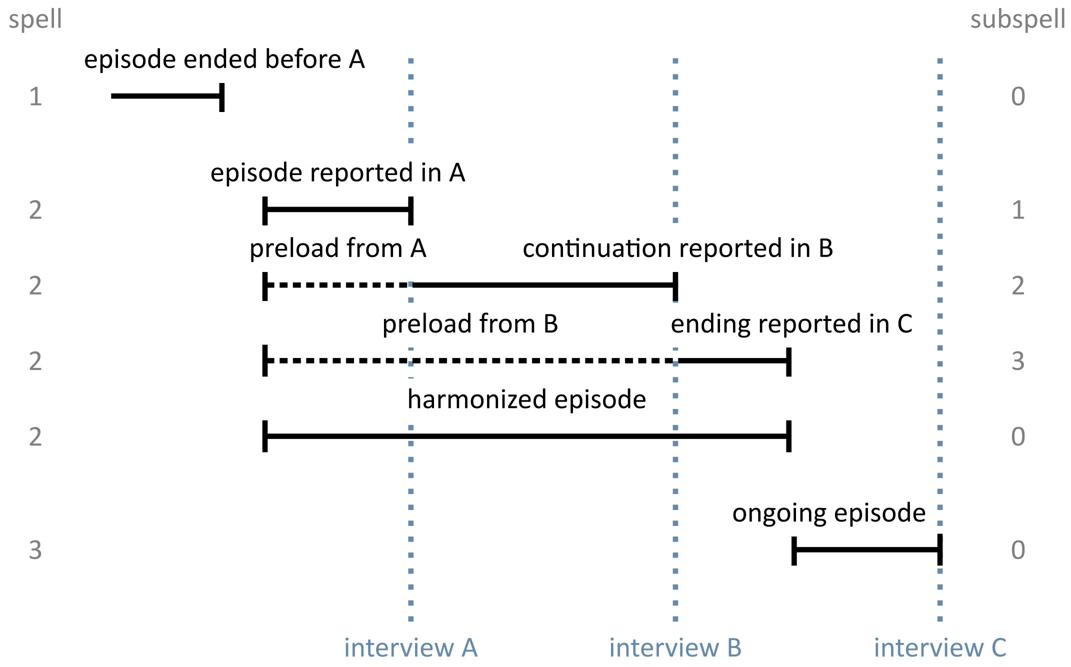

Figure 18: Logic of subspells

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 18).

The sample episodes from Figure 18 correspond to the data structure presented in Table 7 *before* any episode harmonization.[5] 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.

[5] 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).

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

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

For episodes that span over several survey waves, the same information is not collected in each survey wave. In the wave in which an episode is recorded for the first time, all unchanging core information about it is captured. In the example of training episodes, this includes the start date, the type of training (e. g., vocational training or study), the exact name of the training occupation and some other parameters that distinguish this training from others. In later survey waves, this information is no longer requested when updating this episode. However, 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 subepisodes (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. In addition to the data rows for the sub-episodes, there is one data row that provides a summary 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.[6] In case that there are such missings in the first subspell, the next non-missing value from the subsequent subspells is taken instead.

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.[7] 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.[8] 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.

[6] If the missing code -53 (anonymized) is given in the first subspell, this value is copied to the harmonized spell.

[7] If the missing code -53 (anonymized) is given in the last subspell, this value is copied to the harmonized spell.

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

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.

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 review 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. This means, that only not contradicted subspells 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`).[9] As described above, this step is already included in the process of harmonizing episodes.

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

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

[9] 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`).

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 19 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 SC2
** version of this Scientific Use File
global version 11-0-0
** path where the data can be found on your local computer
global datapath Z:/Data/${cohort}/${version}
```

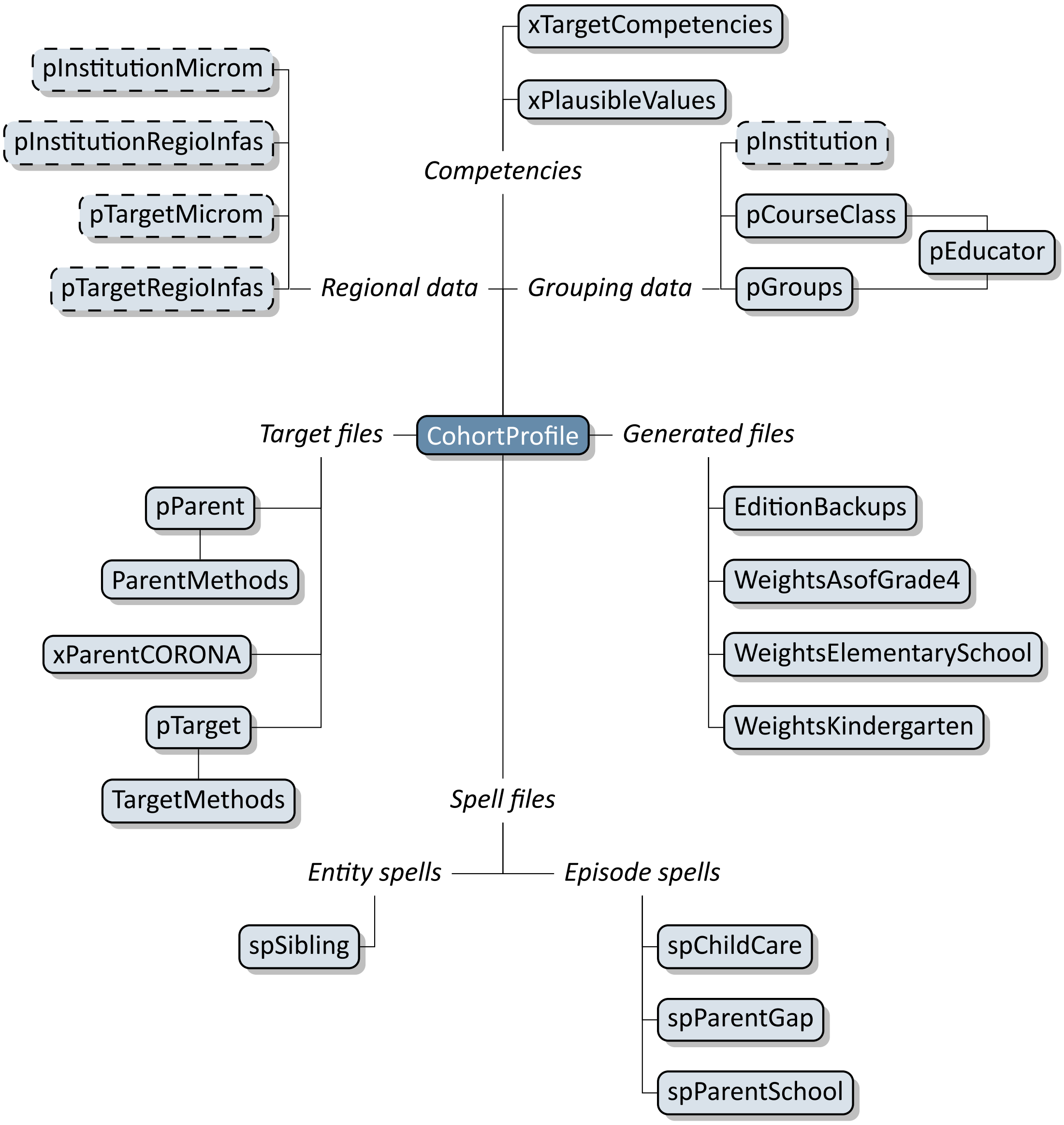

Figure 19: Graphical overview of all data files. Each node represents one data file. Relations are indicated by connection lines. Files with a dashed border are not available in the Download version of the Scientific Use File. Click on a data file to get more information.

4.5.1 CohortProfile

« go back to overview

Description

Paradata on the cohort's panel sample

File structure

long format: 1 row = 1 respondent in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_cc ID_cg ID_cm ID_tg ID_i

Number of variables / number of rows in file

41 / 90,025

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	wave	tx80220	tx80521	tx80522	tx8610y	tx80524
2000727	5	Participation	yes	yes	2014	yes
3004922	5	Participation	yes	yes	2016	yes
3006538	6	Participation	yes	yes	2015	yes
3006930	5	Participation	yes	yes	2014	yes
3018716	6	Participation	yes	yes	2015	yes

The file `CohortProfile` contains all target children of the panel sample. These are all persons with an initial agreement to participation. For each respondent in each wave, the `CohortProfile` contains all ID variables related to this person (from personal ID `ID_t` to kita or school ID `ID_i`), but also meta information like various variables indicating participation e. g., (`tx80220`), sample group (`tx80115`), or availability of specific data (e. g., `tx80522`). In addition, there are variables of the dates when the competence tests (`tx8610`/`tx8611`) took place.

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

Stata 1: Working with CohortProfile

```
** open the data file
use ${datapath}/SC2_CohortProfile_D_${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
```

4.5.2 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 ID_e ID_i wave splink subspell

Other ID variables useful for linkage

mergevars

Number of variables / number of rows in file

15 / 204

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

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
2001310	10	pTarget	tm00055	ID_t wave tx20100	7.00	1.00
3017552	10	pTarget	tm00055	ID_t wave tx20100	7.00	2.00
3018067	10	pTarget	tm00055	ID_t wave tx20100	7.00	1.00
3018447	10	pTarget	tm00055	ID_t wave tx20100	7.00	1.00
3018803	10	pTarget	tm00055	ID_t wave tx20100	7.00	1.00

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 2: Working with EditionBackups

```
** In this example, we want to restore the original values in the variable
** tm00055 (Learning materials) of datafile pTarget

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

** only keep rows containing data of the aforesaid variable
keep if dataset=="pTarget" & varname=="tm00055"

** check which variables we need for merging
tab mergevars

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

** rename the variables to emphasize affiliation
rename sourcevalue_num tm00055_source
rename editvalue_num tm00055_edit

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

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

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

** check all edition made
list ID_t wave tm00055* if _merge==3, nolab

** replace the variable in the datafile with its original value
replace tm00055=tm00055_source if _merge==3
```

4.5.3 ParentMethods

« go back to overview

Description

Paradata from the parents interview

File structure

long format: 1 row = 1 parent in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_int

Number of variables / number of rows in file

35 / 56,469

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	wave	px80209	ID_int	px80301	px80302
2001008	3	25.89167	1816	[w] female	50-65 years
2001254	2	27.73333	1830	[w] female	up to 29 years
2001394	4	27.64833	2214	[w] female	up to 29 years
3007627	3	26.96667	2018	[w] female	30-49 years
3007753	3	28.10833	2015	[m] male	up to 29 years

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

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

Stata 3: Working with ParentMethods

```
** open the data file
use ${datapath}/SC2_ParentMethods_D_${version}.dta, clear

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

** check out response code by wave
tab wave px80207

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

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

4.5.4 pCourseClass

« go back to overview

Description

Data about the class background

File structure

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

ID variables needed to identify a single row

ID_cc ID_e wave ex20100

Other ID variables useful for linkage

Number of variables / number of rows in file

361 / 2,916

Contains data from waves

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

Exemplary variables

ID_cc	Course-ID: Grade
wave	Wave
ID_e	ID teacher/educator
e451010	Class: number of students with a migrant background (approximately)
e79202a_D	Students with at least one parent with a higher education degree (in %)
e227400_R	Class: number of female students
e227401_R	Class: number of male students
e22740a	Class: teacher assessment: interest

Exemplary data snapshot

ID_cc	wave	ID_e	e227400_R	e227401_R
10026330001	5	1018905	8	12
1002614101	4	1011576	10	12
1002922101	3	1012312	12	13
10027670001	5	1018862	14	8
1002821102	4	1012072	6	14

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

In some cases, more than one teacher reported information about a single class, although this was not intended by the survey design. In such cases, we made a suggestion which data to use in variable ex20100.[10]

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

[10] The data row with the least missing values is being suggested.

Stata 4: Working with pCourseClass

```
** there are 2 scenarios for using pCourseClass
* 1) merging detailed teacher information from pEducator via ID_e:

** Opening pEducator and dropping not recommended observations
use ${datapath}/SC2_pEducator_D_${version}.dta, clear
keep if ex20100 == 1

** saving result in a temporary dataset
tempfile educator
save `educator'

** Opening pCourseClass and dropping not recommended observations
use ${datapath}/SC2_pCourseClass_D_${version}.dta, clear
keep if ex20100 == 1

** merging the previously stored temporary dataset
merge 1:1 ID_e wave using `educator', keep(master match) keepusing(e762110) nogen

** inspect the teachers sex
tab e762110


* 2) mergeing class information to CohortProfile in order to merge additional panel
  information
** opening and preparing pCourseClass
use ${datapath}/SC2_pCourseClass_D_${version}.dta, clear

** since there may be more than one class teacher, the dataset must first be limited
  to one observation per class by removing duplicates
duplicates tag ID_cc wave, gen(dups)
bysort ID_cc dups: drop if dups ==1 & _n!=1             // keep only first
  observation

** saving result in a temporary dataset
tempfile class
save `class'

** Opening CohortProfile
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear

** remove all observations without class information - can be added again in a later
  step
drop if inlist(ID_cc,-54,-55)

** merge previously stored pCourseClass information
merge m:1 ID_cc wave using `class', assert(master match) keepusing(e451000_D) nogen

** inspect number of students with a migrant background (in %)
summarize e451000_D if e451000_D>0
```

4.5.5 pEducator

« go back to overview

Description

Personal information about educators and teachers

File structure

long format: 1 row = 1 educator in 1 wave

ID variables needed to identify a single row

ID_e wave ex20100

Other ID variables useful for linkage

ID_cc

Number of variables / number of rows in file

157 / 4,722

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_e	wave	e400000	e41100a_g1	e537090	e537150_R	e762110
1018954	5	3	1	1	1996	[w] female
1011649	4	3	1	1	2002	[w] female
1019753	6	3	1	1	1997	[w] female
1012859	5	3	1	1	1994	[w] female
1018919	5	2	1	1	2007	[w] female

Kindergarten educators and school teachers were interviewed as context persons by PAPI questionnaires. This data is made available in the file pEducator. The scope of information comprises various personal attributes of the persons, e. g., gender (e762110), year of birth (e76212y_D), or migration background (e400000), as well as attitudes such as aspects of career choice (e g., e536031).

This file contains all educators from the sample, no matter if they were kindergarten educators or class teachers of the target children. Please note that there is no direct link between the children and the educators made available. This is due the following reasons:

- Due to missing detail in instructions, in some cases more than one educator or teacher answered the survey (see variable ex20100 in, e. g., pCourseClass).
- There is no natural relationship in the survey design between educators and children. The educators have only been interviewed about themselves, as well as about the group or class.

To map those facts to the data, the educator was only attached to the specific group or class (pCourseClass, pGroups), but not to the children directly (i. e., there is no variable ID_e in CohortProfile). See the example below on how you are supposed to use the data from pEducator.

Stata 5: Working with pEducator

```
** open the pEducator dataset
use ${datapath}/SC2_pEducator_D_${version}.dta, clear

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

** note that we have data from both Educators (wave 1-3) and Teachers (wave 4-6)
** in this file
tab wave

** save temporarily
tempfile educators
save `educators'

** we can now merge them to both pGroups as well as pCourseClass:

use ${datapath}/SC2_pGroups_D_${version}.dta, clear
merge m:1 ID_e wave using `educators', keep(master match) nogen


use ${datapath}/SC2_pCourseClass_D_${version}.dta, clear
merge m:1 ID_e wave using `educators', keep(master match) nogen
```

4.5.6 pGroups

« go back to overview

Description

Data about the kindergarten group background

File structure

long format: 1 row = 1 kindergarten group in 1 wave

ID variables needed to identify a single row

ID_group wave ex20100

Other ID variables useful for linkage

ID_e

Number of variables / number of rows in file

145 / 2,772

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_e	ID teacher/educator
ID_group	ID Kindergarten group
wave	Wave
ex20100	Data line recommended for linkage
e21141e	Kindergarten: frequency of visits: forest, park etc.
e21951a	Kindergarten: availability: picture books
e21951d	Kindergarten: availability: dolls
e21951g	Kindergarten: availability: music instruments
e21140d	Kindergarten: frequency of activity: puzzles and the like
e21140h	Kindergarten: frequency of activity: sport and the like

Exemplary data snapshot

ID_e	ID_group	wave	ex20100	e21141e	e21951d
1004114	1000801101	1	1	4	About half of all children
1004264	1000856101	1	1	5	Some children
1004161	1000762102	1	1	3	Almost all children
1003755	1000705101	1	1	5	Some children
1004964	1001164103	1	1	4	Some children

This data file contains all the information surveyed from the educator about the kindergarten group. This is for example the number/percentage of girls (e217401_R), boys (e217402_D), children in total by birth year (e. g. e217411_R), and children with a migration background by birth year (e. g. e45110a_R) or equipment of the kindergarten (e. g. picture books e21951a). The educator reporting this information can be identified via ID_e.

In some cases, more than one educator reported information about a single kindergarten group, although this was not intended by the survey design. In such cases, we made a suggestion which data to use in variable ex20100.[11]

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

[11] The data row with the least missing values is being suggested.

Stata 6: Working with pGroups

```
** open the data file
use ${datapath}/SC2_pGroups_D_${version}.dta, clear
label language en

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

** temporarily save data
tempfile groups
save `groups'

** open CohortProfile
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear
label language en

** merge previously created dataset
merge m:1 ID_group wave using `groups', keep(master match) nogen

** tabulate 'state of school (west/east)' (from CohortProfile) against 'frequency of
  visits: museum' (from pGroups)
tab e21141a tx80109_g1
```

4.5.7 pInstitution

« go back to overview

Description

Context data collected from the head of institution about the kindergarten or school

File structure

long format: 1 row = 1 institution in 1 wave

ID variables needed to identify a single row

ID_i wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

712 / 1,655

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_i	wave	h190011	h22202a	h227000	h229000	h535010
1002747	3	1	-54	6	1	1
1002775	3	25	-54	22	1	4
1002887	3	22	-54	19	1	1
1002574	3	8	-54	13	1	3
1002848	3	6	-54	22	1	2

Data about the kindergarten or school as institutional context has been surveyed from the kindergarten management or school principal via PAPI mode. The resulting data file pInstitution) contains this information, including data like sponsor (h219005) or administration h229000, the total number of children by birth year (e. g., h217012_w1) or students at the school (h227100), as well as background on the infrastructure (e. g., schools within a radius of 10km in h535010), and more.

Please note that this datafile is only available in the RemoteNEPS version!

Stata 7: Working with pInstitution

```
** open the CohortProfile
use ${datapath}/SC2_CohortProfile_R_${version}.dta, clear

** merge the size of the school to CohortProfile using school ID
merge m:1 ID_i wave using ${datapath}/SC2_pInstitution_R_${version}.dta, ///
  keepusing(h227100) nogen assert(master match)

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

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

tab size
```

4.5.8 pInstitutionMicrom

« go back to overview

Description

regional data about the geographical area of institution

File structure

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

ID variables needed to identify a single row

ID_i wave regio

Other ID variables useful for linkage

ID_regio

Number of variables / number of rows in file

188 / 4,298

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_i	wave	regio	ID_regio	mso_k_ausland	mbe_k_haustyp	mpi_k_dichte
1002841	3	1	108353	2	2	7
1002918	3	1	113013	8	2	6
1002889	3	1	144453	4	1	9
1002739	4	1	101094	4	1	8
1002605	4	1	155578	2	3	3

The data file `pInstitutionMicrom` is only available **On-site**. You cannot work with this file having only access to the Download or Remote data version.

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

All those levels are available for every institution and every wave. There is a lot of regional information in this file, including percentage of foreigners, unemployment rate, family structure, milieu types, car type/density, insurances, only to name a few. To clarify this, those details are **not** about the respondents or the institutions but about the regional level (e. g., the unemployment rate is not the rate at the institution but the rate in the municipality the institution resides). Please be aware that there is a complete documentation about this data file that not only lists all variables but also has a description of the background. See Section 1.2 on how to find this document.

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

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

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

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

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

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

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

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

4.5.9 pInstitutionRegioInfas

« go back to overview

Description

regional data about the geographical area of institution

File structure

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

ID variables needed to identify a single row

ID_i wave regio

Other ID variables useful for linkage

none

Number of variables / number of rows in file

68 / 1,116

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_i	regio	tx44288	tx44294	tx44298	tx44001	tx44318
1000803	3	12.9	19855	101.37	2.24	35.00
1000893	2	10.5	17603	977.70	1.74	53.42
1000790	2	16.4	18573	91.67	1.89	45.30
1000873	2	17.3	16512	6.56	2.55	22.92
1000906	1	16.4	16908	19.76	2.13	36.75

The data file `pInstitutionRegioInfas` is only available **On-site**. You cannot work with this file having only access to the Download or Remote data version.

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

There is a lot of regional information in this file, including purchasing power per resident in EUR (`tx44294`), companies in total per km^2 (`tx44298`), residents per household (`tx44001`), and so on. As in `pInstitutionMicrom`, those details are **not** about the respondents but about the regional level (e. g., the unemployment rate is not the rate at the institution but the rate in this municipality). Please be aware that there is a complete documentation about this data file that not only lists all variables but also has a description of the background. See Section 1.2 on how to find this document.

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

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

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

** existing regional levels are:
tab regio

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

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

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

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

4.5.10 pParent

« go back to overview

Description

Data surveyed from parents

File structure

long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

1,021 / 38,865

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave	Wave
p731905	Professional position respondent
p731955	Professional position partner
p731701	Relationship Respondent to target child
p741001	Household size
p400500_g1	Generation status
p743040	TC in HH
p731905	Professional position respondent
p73170y	Date of birth respondent: year
p401100	German citizenship respondent
p731116	Gender partner

Exemplary data snapshot

ID_t	wave	p731905	p731955	p741001	p400500_g1	p743040
3004439	3	1	1	5	3	yes
3005035	4	2	2	5	6	yes
3006563	3	2	2	4	6	yes
3017226	3	1	1	4	3	yes
3019369	3	2	2	4	8	yes

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

Stata 10: Working with pParent

```
** open the CohortProfile
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear

** merge occupation of parents (both respondent and partner) from pParent
merge 1:1 ID_t wave using ${datapath}/SC2_pParent_D_${version}.dta, ///
  keepusing(p731905 p731955) nogen assert(master match)

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

** recode missings
nepsmiss p731905 p731955

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

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

** check the distribution of parents occupation in current type of school
tab2 p731905 p731955 tx80106
```

4.5.11 pTarget

« go back to overview

Description

Data surveyed from and about the target children

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

tx20100 ID_e

Number of variables / number of rows in file

695 / 47,818

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave	Wave
t514001	Satisfaction with life
t400000_g1D	Country of birth (Germany/abroad)
t41000a_g1	Mother tongue (number references)
t514004	Satisfaction with family life
t66002a_g1	Self concept school
t66206f_g1	IILS-C: Conventional interests
e66800a_g1	Big Five: extraversion
e41271a	Language support: individual support
t34001a	Reading – normal school day

Exemplary data snapshot

```
     ID_t           wave        t514001            t400000_g1D      t41000a_g1      t514004
  3005235              7              9      missing by design             -54           10
  3005533              8              9      missing by design             -54           10
  3007365             11              9      missing by design             -54           10
  3018814             10              8      missing by design             -54           10
  3019206              7             10      missing by design             -54           10
```

The file pTarget contains all information surveyed from the target persons (children) themselves and from educators about the specific child. A lot of items are included in this data file, ranging from demographic topics (e. g., country of birth t400000_g2R) to attitudes or expectations (e. g., self-concept German t66000a_g1, activities (e. g., sports t262000_g1), network issues (e. g., share friends with migrant background t451200), and many more.

Stata 11: Working with pTarget

```
** open the CohortProfile
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear

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

** merge country of birth from pTarget
merge 1:1 ID_t wave tx20100 using ${datapath}/SC2_pTarget_D_${version}.dta, ///
  keepusing(t400000_g1D) nogen

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

** recode missings
nepsmiss t400000_g1D

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

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

** check the above alteration
tab t400000_g1D wave, miss

** check the distribution between migration and current type of school
tab tx80106 t400000_g1D
```

4.5.12 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 / 94,190

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

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
2000996	4	1	150659	9	4	1
2002216	4	1	155375	1	3	9
2002264	4	1	112273	6	4	8
2003583	3	1	158330	9	6	2
3017582	4	1	113509	5	2	7

The data file `pTargetMicrom` is only available **On-site**. You cannot work with this file having only access to the Download or Remote data version.

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

Numerous regional indicators are provided, e. g. the percentage of foreigners, unemployment rate, family and age structure, milieu types, car type density, distribution of insurances, etc. To clarify, this information does **not** refer to individuals, but to regional units to which respondents belong via their place of residence. Accordingly, the unemployment rate, for example, indicates the proportion of unemployed people in the population of a given region.

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

Stata 12: 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
```

4.5.13 pTargetRegioInfas

« go back to overview

Description

Small-scale regional indicators on respondents' place of residence

File structure

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

ID variables needed to identify a single row

ID_t wave regio

Other ID variables useful for linkage

none

Number of variables / number of rows in file

68 / 11,380

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	regio	tx44294	tx44298	tx44001
2003072	1	19745	25.79	2.27
2003595	1	17886	192.09	1.77
2000770	3	20179	14.71	2.34
2003213	3	17709	92.69	1.84
2000983	2	17606	9.29	2.27

The data file pTargetRegioInfas is only available **On-site**. You cannot work with this file having only access to the Download or Remote data version.

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

As in pTargetMicrom these data do **not** refer to the respondents themselves, but to the regional levels in which the respondents live (i. e., the unemployment rate, for example, indicates the proportion of unemployed people in the population of a given region such as the municipality).

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 13: Working with pTargetRegioInfas (find R example here)

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

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

** existing regional levels are:
tab regio

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


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

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

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

4.5.14 spChildCare

« go back to overview

Description

Spell data on child care episodes relating to the target child

File structure

spell format: 1 row = 1 episode of 1 respondent

ID variables needed to identify a single row

ID_t spell sptype

Other ID variables useful for linkage

wave

Number of variables / number of rows in file

12 / 11,393

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave	Wave
sptype	Care: type of episode
spell	Spell number
startm	Care: start (month)
starty	Care: start (year)
endm	Care: end (month)
endy	Care: end (year)
timepweek	Care: duration per week (hours)
fee_R	Care: fees (euros)
fee_D	Care: fees (euros, categorized)
pb10110	Care: lunch included in fee

Exemplary data snapshot

ID_t	wave	sptype	spell	startm	starty	endm	endy
2000700	1	1	1	August	2008	May	2011
2001900	1	1	1	November	2007	May	2011
2002189	1	2	1	May	2006	August	2006
2002949	1	6	1	February	2007	May	2011
2003233	1	6	1	June	2007	July	2009

The data file spChildCare contains all child care episodes relating to the target child, differentiated according to the carer (e. g., grandparent, nanny, childminder); see the variable sptype. Besides the start and end dates of the respective episodes (startm/y, endm/y), it essentially contains structural information such as duration of care (timepweek) or fees (fee_R/D).

Stata 14: Working with spChildCare

```
** open the data file
use ${datapath}/SC2_spChildCare_D_${version}.dta, clear
label language en

** check who provided the child care
tab sptype

** only keep episodes where child care has been provided by nanny
keep if sptype==4
```

```
** generate the total duration of the episode (in months)
generate ep_start=ym(starty, startm)
generate ep_end=ym(endy, endm)
generate duration=ep_end-ep_start+1

** check if this was correctly computed
list startm starty endm endy ep_start ep_end duration in 1/10

** display basic statistics for the duration of nanny child care
summarize duration
```

4.5.15 spParentGap

« go back to overview

Description

Gap episodes reported by the parents

File structure

spell format: 1 row = 1 gap of 1 respondent

ID variables needed to identify a single row

ID_t spell

Other ID variables useful for linkage

wave splink spms

Number of variables / number of rows in file

17 / 301

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	wave	ps2911m	ps2911y	ps2912m	ps2912y	ps2912c
2001188	7	8	2016	9	2016	.
2001199	6	2	2016	4	2016	1
2002046	9	9	2017	5	2019	1
3005325	5	11	2013	5	2015	1
3008520	7	7	2016	9	2016	.

The datafile spParentGap contains information on biographical gaps of the target persons (**reported by the parents**). Note that these are not gaps in the lifecourse of the parent, but of the children. The spells in this file refer to different types of gaps that can be distinguished by the variable ps29101. Further details on single gaps refer to the start and end date (ps2911m/y, ps2912m/y) as well as to the ongoing of the gap episode (p2912c)

Stata 15: Working with spParentGap

```
** open the Gap data file
use ${datapath}/SC2_spParentGap_D_${version}.dta, clear
label language en

** get an overview about the type of gaps
tab ps29101
```

4.5.16 spParentSchool

« go back to overview

Description

General schooling history reported by the parents

File structure

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

ID variables needed to identify a single row

ID_t spell subspell

Other ID variables useful for linkage

wave splink

Number of variables / number of rows in file

41 / 38,714

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	subspell	spell	wave	p723020	p72302m	p72302y	p723080
3005505	1	1	3	1	6	2013	1
3006445	1	1	3	1	6	2013	1
3017483	1	1	3	1	8	2013	1
3018595	1	1	3	1	8	2013	1
3019005	1	1	3	1	7	2013	14

This dataset covers each child's general education history starting from school entry (**reported by the parents**). The file covers the start and end dates of school episodes (`p72301m/y`, `p72302m/y`) as well as several information about the respective episode (e. g., type of school `p723080`, school authority `p723180`, reason for school change `p723130`).

A new episode is generated only if the school type changes. That is, a change from one elementary school 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 16: Working with spParentSchool

```
** open the data file
use ${datapath}/SC2_spParentSchool_D_${version}.dta, clear

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


** evaluate how many children have school episodes already
distinct ID_t

** check the distribution of the number of episodes per child
summarize spell

** generate an indicator if a child ever visited a public school (vs. church/private
  schools)
bysort ID_t: egen public =  max(p723180==1)

** create minimal dataset
keep ID_t public
duplicates drop
tempfile tmp
save `tmp'

** open the CohortProfile data file
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear

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

** you now have an enhanced version of CohortProfile, enriched by
** information from the spell module.
```

4.5.17 spSibling

« go back to overview

Description

Siblings of the respondent

File structure

entity format: 1 row = 1 sibling of 1 respondent

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

37 / 11,927

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

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

Exemplary data snapshot

ID_t	wave	p73221y	p732220	p732230
2002786	2	1995	1	half brother/half sister
2003181	2	2002	2	biological brother/biological sister
3007119	4	2003	2	biological brother/biological sister
3017426	4	2002	2	biological brother/biological sister
3017927	4	2013	1	half brother/half sister

The file spSibling contains information on all siblings of the respondent (**reported by the parent**). Each sibling is stored in one separate row, providing details about the date of birth (p73221m/y), gender (p732220), employment status (p732401), and highest school-leaving qualification (p732313).

Stata 17: 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
use ID_t tx8050m tx8050y using ${datapath}/SC2_CohortProfile_D_${version}.dta, clear

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

label language en
tempfile temp
save `temp'

** now, open the spSibling data file
use ${datapath}/SC2_spSibling_D_${version}.dta, clear
label language en

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

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

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

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

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

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

4.5.18 TargetMethods

« go back to overview

Description

Paradata from the targets interview

File structure

long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_int

Number of variables / number of rows in file

24 / 29,430

Contains data from waves

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

Exemplary variables

ID_t	ID target
wave	Wave
cohort	NEPS Starting Cohort
tx80201	Interview: survey mode (start)
tx80202	Interview: survey mode (realized case)
tx80221	Interview: evaluable data set?
tx80301	Interviewer: gender
tx80303	Interviewer: highest school-leaving qualification
ID_int	Interviewer: ID
tx80205	Interview: interview interrupted
tx80200	Interview: number of all contact attempts

Exemplary data snapshot

ID_t	wave	tx80201	tx80221	tx80301	tx80303	ID_int
2002133	6	CAPI	does apply	2	2	1874
2003233	6	CAPI	does apply	1	7	1860
3006300	9	CAPI	does apply	2	17	2572
3006993	9	CAPI	does apply	2	7	2347
3017788	9	CAPI	does apply	1	7	1329

This dataset offers a variety of information on the data collection via personal interviews with the target children, e. g., gender (tx80301) and age (tx80302) of the interviewer, interview duration (tx80209), response code (tx80207), and survey mode (tx80202).

This file contains all contacted respondents whether an interview was realized or not (see variable tx80207 for more details). Thus, TargetMethods includes more cases than the dataset pTarget.

Stata 18: Working with TargetMethods

```
** open the data file
use ${datapath}/SC2_TargetMethods_D_${version}.dta, clear

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

** check out response code by wave
tab wave tx80207

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

** get an overview on the count of contact attempts
summarize tx80200
```

4.5.19 WeightsAsofGrade4

« go back to overview

Description

Weights for sample groups 1, 2, 3

File structure

wide format: 1 row = 1 target

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

22 / 9,337

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
w_p6	Panel input weight for wave 6 / grade 4 (calibrated)
w_p6_joint	Total panel input weight for wave 6 / grade 4 (calibrated)
w_t6	Weight wave 6
w_tp6	Weight for joint participation of children and parents in wave 6
w_t7	Weight wave 7
w_tp7	Weight for joint participation of children and parents in wave 7

Exemplary data snapshot

ID_t	w_p6	w_t6	w_tp6	w_t7	w_tp7
3017825	76.52454	0.66837	0.56509	0.51693	0.36529
2001929	76.10030	0.66992	0.58073	0.49572	0.32451
3007729	362.75215	3.14662	2.74534	2.40998	1.77468
3006687	50.19899	0.43005	0.37004	0.30167	0.23921
3004527	37.17259	0.32180	0.28133	0.26792	0.18186

Due to the particular survey design (see Section 2.2), the NEPS provides various kinds of weights for kindergarten children and elementary school students together with design information.

- **WeightsKindergarten**: for kindergarten children (groups 1 and 3, frozen in wave 6 and not continued)
- **WeightsElementarySchool**: for elementary school students (groups 1 and 2)
- **WeightsAsofGrade4**: for all grade 4 students transferring to lower secondary education (groups 1 to 3)

The weighting dataset referring to kindergarten children provides all cross-sectional and longitudinal weights in a trimmed and standardized form. The weighting dataset referring to elementary school students and the dataset with joint weights for all targets as of grade 4 provide all cross-sectional and longitudinal weights standardized.

4.5.20 WeightsElementarySchool

« go back to overview

Description

Weights for sample groups 1, 2

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

43 / 9,337

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
ID_i	Institution ID
tstud_st	Substudy of first survey
group	Grouping variable for children in Kindergarten and school context
stratum_imp4_R	implicit stratum 4 (organizing institution)
w_i	design weight for institutions
w_t	Panel input weight for children with parental consent
w_t3	Cross-sectional weight for targets participating in wave 3

Exemplary data snapshot

ID_t	ID_i	group	w_i	w_t	w_t3
3006071	1002768	1	14.96117	35.48356	0.36958
3008254	1002710	1	17.34366	24.11441	0.25130
3005191	1002759	1	25.35629	46.09499	0.54615
3004922	1002753	1	18.20746	46.48638	0.48392
3005340	1002743	1	35.02740	50.78750	0.52966

Due to the particular survey design (see Section 2.2), the NEPS provides various kinds of weights for kindergarten children and elementary school students together with design information.

- **WeightsKindergarten**: for kindergarten children (groups 1 and 3, frozen in wave 6 and not continued)
- **WeightsElementarySchool**: for elementary school students (groups 1 and 2)
- **WeightsAsofGrade4**: for all grade 4 students transferring to lower secondary education (groups 1 to 3)

The weighting dataset referring to kindergarten children provides all cross-sectional and longitudinal weights in a trimmed and standardized form. The weighting dataset referring to elementary school students and the dataset with joint weights for all targets as of grade 4 provide all cross-sectional and longitudinal weights standardized.

4.5.21 WeightsKindergarten

« go back to overview

Description

Weights for sample groups 1, 3

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

20 / 2,996

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
ID_i	Institution ID
tstud_st	Substudy of first survey
group	Grouping variable for children in Kindergarten and school context
w_i	design weight for institutions
w_t	Panel input weight for children with parental consent
w_t1	Cross-sectional weight for targets participating in wave 1
w_tp1	Cross-sectional weight for joint participation in wave 1
w_t2	Cross-sectional weight for targets participating in wave 2

Exemplary data snapshot

ID_t	ID_i	group	w_i	w_t	w_t1	w_t2
2001078	1000821	2	278.27684	407.87282	1.95956	1.84049
2001688	1000731	2	79.23402	95.46125	0.45863	0.43076
2002502	1000822	2	366.19323	1026.84184	4.26743	4.45437
2000572	1000900	2	74.47998	59.31077	0.28495	0.26763
2001308	1000812	2	163.33329	142.55915	0.68490	0.64328

Due to the particular survey design (see Section 2.2), the NEPS provides various kinds of weights for kindergarten children and elementary school students together with design information.

- **WeightsKindergarten**: for kindergarten children (groups 1 and 3, frozen in wave 6 and not continued)
- **WeightsElementarySchool**: for elementary school students (groups 1 and 2)
- **WeightsAsofGrade4**: for all grade 4 students transferring to lower secondary education (groups 1 to 3)

The weighting dataset referring to kindergarten children provides all cross-sectional and longitudinal weights in a trimmed and standardized form. The weighting dataset referring to elementary school students and the dataset with joint weights for all targets as of grade 4 provide all cross-sectional and longitudinal weights standardized.

4.5.22 xParentCORONA

« go back to overview

Description

Data collected in May 2020 regarding the impact of the corona pandemic

File structure

wide format: 1 row = 1 target

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

179 / 1,587

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave	Wave
p515051	Willingness to take risks in general
pm00015	Systemically important profession
pm00016	Change working time
pm00017	Change work place
pm00018	Change status
pm00019	Support Employer
pm00020	Change order situation
pm00051	Homeschooling, equipment

Exemplary data snapshot

```
    ID_t   wave   p515051                              pm00018                pm00019
 2000727      .         7  Instructed reduction of vacation/overtime  reasonably well supported
 3006216      .         5                          short-time work         very well supported
 3017726      .         4                          short-time work              well supported
 3018254      .         7                          short-time work  reasonably well supported
 3018599      .         2                          short-time work         very well supported
```

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.

- 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 be able to survey these data in a timely manner, the 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 a regular survey wave, the information is marked with a missing wave indicator (`wave==.`) in the data and separately stored in this file.

In Starting Cohort 2, the parents served as informants, but most of the information refer to the target children. Selected questions have then been integrated in an additional module on the Corona pandemic, which became part of the subsequent main surveys in all starting cohorts. These data are integrated in the file `pTarget`.

Stata 19: Working with xParentCORONA

```
** open the file xParentCORONA
use ${datapath}/SC2_xParentCORONA_D_${version}.dta, clear

** note that this is an 'intermediate' wave
tab wave, nolab

** open the pParent file
use ${datapath}/SC2_pParent_D_${version}.dta, clear

** add the information from xParentCORONA
append using ${datapath}/SC2_xParentCORONA_D_${version}.dta

** note that the identificators did not change
isid ID_t wave

** we now have an additional wave in this file
tab wave
```

4.5.23 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

315 / 9,313

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2011)
wave_w3	Row contains data from wave 3 (2013)
vok1_pv1	Vocabulary: cross-sectional plausible value 1
vok1_pv2	Vocabulary: cross-sectional plausible value 2
mag1_pv1	Math: cross-sectional plausible value 1
mag1_pv1u	Math: longitudinal plausible value 1
mag1_pv2u	Math: longitudinal plausible value 2
mag1_pv10u	Math: longitudinal plausible value 10

Exemplary data snapshot

ID_t	wave_w1	wave_w3	vok1_pv1	vok1_pv2	mag1_pv1	mag1_pv1u
2001482	1	1	0.94599	1.72785	1.79409	1.49816
2000977	1	1	0.95670	0.94384	1.06734	2.22496
2000684	1	1	0.94599	0.70021	1.66183	2.24064
2002120	1	1	0.77257	1.11941	0.61958	2.39229
2001270	1	1	0.62865	1.06844	0.33453	1.15738

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), PV are suitable for more precise inferential statistical tests in correlation and mean value analyses.

PV 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.

See for more information the NEPS Survey Paper (Scharl et al., 2020) and our website:

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

Stata 20: 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!
```

4.5.24 xTargetCompetencies

« go back to overview

Description

Competence 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

1,343 / 9,317

Contains data from waves

1 2 3 4 5 6 7 8 9 10 11

Exemplary variables

ID_t	ID target
wave_w1	Row contains data from wave 1 (2011)
wave_w5	Row contains data from wave 5 (2014/2015)
vok1_sc3	Vocabulary: sum
vok1_sc1u	Vocabulary: WLE (uncorrected)
sck1_sc1	Scientific literacy: WLE (corrected)
sck1_sc2	Scientific literacy: standard error of WLE (corrected)
dsk2_sc3	Digit span: sum
icg3_sc1	ICT literacy: WLE (corrected)
org4_sc1b	Orthography (set 2): WLE (corrected)

Exemplary data snapshot

ID_t	wave_w1	wave_w5	vok1_sc3	sck1_sc1	dsk2_sc3
2000805	1	1	63	1.64001	6
2001579	1	1	57	0.08262	6
2002375	1	1	52	0.64726	6
2002620	1	1	51	0.08262	6
2002273	1	1	60	1.22713	8

The file `xTargetCompetencies` contains data from competence assessments in several domains conducted in the kindergarten, the school and at home. Scored single item variables as well as generated scale indices are available in a cross-sectional format. The overview in Table 2 informs about the tested competencies for each survey wave. The variables `wave_w*` can be used to select rows only containing data from a specific wave.

Detailed information about the naming conventions for competence items are provided in Section 3.2.2.

Stata 21: Working with xTargetCompetencies

```
** open datafile
use ${datapath}/SC2_xTargetCompetencies_D_${version}.dta, clear

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

** as the 'x' in the filename indicates, this is a cross sectional file
** (no wave structure). You can verify this by asking if one row is
** solely identified by the respondents ID
isid ID_t

** note that competence testing has been conducted in multiple waves
** an indicator marks if a row contains information for a specific wave
tab1 wave_w*

** to work with competence data, you might want to merge it to CohortProfile.
** if you want to keep the panel logic (and not only add all competencies
** to every wave), you need a mergeable wave variable in xTargetCompetencies.
** in this example, we focus on math competencies, which have been tested in wave 1.
generate wave=1

** now, remove cases which did not took part in the testing
drop if wave_w1==0

** and reduce the dataset to the relevant variables
keep ID_t wave mag9_sc1 mag9_sc2

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

** and merge this to CohortProfile
use ${datapath}/SC2_CohortProfile_D_${version}.dta, clear
merge 1:1 ID_t wave using `tmp', nogen
```

A References

Aßmann, C., Steinhauer, H. W., Würbach, A., Zinn, S., Hammon, A., Kiesl, H., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2019). Sampling designs of the National Educational Panel Study: Setup and panel development. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 35–55). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0_3

Berendes, K., Linberg, T., Müller, D., Wenz, S., Roßbach, H.-G., Schneider, T., & Weinert, S. (2019). Kindergarten and elementary school: Starting Cohort 2 of the National Educational Panel Study. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 215–230). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-658-23162-0

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*.

FDZ-LIfBi. (2024). *Data Manual NEPS Starting Cohort 2–Kindergarten, From Kindergarten to Elementary School, Scientific Use File Version 11.0.0*. 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.

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. (2024-a). *National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC2:11.0.0.

NEPS Network. (2024-b). *Starting Cohort 2: Kindergarten (SC2), Wave 11, Questionnaires (SUF Version 11.0.0)*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pelz, S. (2023). *NEPS Technical Report: Implementation of the ISCED-97, CASMIN and Years of Education Classification Schemes in SUF Starting Cohort 2*. 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.

Steinhauer, H. W., & Zinn, S. (2016). *NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 3 of the National Educational Panel Study (Waves 1 to 5)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Wenzig, K. (2012). *NEPS-Daten mit DOIs referenzieren* (RatSWD Working Paper Series). Rat für Sozial- und Wirtschaftsdaten, Berlin.

Zinn, S., Würbach, A., Steinhauer, H. W., & Hammon, A. (2018). *Attrition and selectivity of the NEPS Starting Cohorts: An overview of the past 8 years* (NEPS Survey Paper No. 34). 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  <- "2"

# version of this Scientific Use File
version <- "11-0-0"
```

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 22: 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 23: Working with pInstitutionMicrom

```
# open pTargetMicrom datafile. Note that this data file is only available OnSite!'
Microm <- read.neps("pInstitutionMicrom")

# additionally to ID_i and wave, line identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(Microm[,c("ID_i", "wave" ,"regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate


# tabulating wave against regio shows availability of all levels
# in wave 5 and 7, but only the most detailed level available
# in wave 1 and 3 (usually housing level)
addmargins(table(Microm$wave, Microm$regio))

# only keep housing level
Microm <- subset(Microm, Microm$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
Microm <- merge(CohortProfile, Microm, by = c("ID_i", "wave"), all = TRUE)
```

R 24: Working with pInstitutionRegioInfas

```
# open  datafile. Note that this data file is only available OnSite!
RegioInfas <- read.neps("pInstitutionRegioInfas")

# identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(RegioInfas[,c("ID_i", "regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate

# existing regional levels are:
table(RegioInfas$regio)

# only keep housing level
RegioInfas = subset(RegioInfas, RegioInfas$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
RegioInfas <- merge(CohortProfile, RegioInfas, by = c("ID_i","wave"), all = TRUE)
```

R 25: 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 26: Working with pTargetRegioInfas

```
# open RegioInfas datafile. Note that this data file is only available OnSite!
RegioInfas <- read.neps("pTargetRegioInfas")

# identification in this file is done
# via variable regio, denoting the regional level of information
anyDuplicated(RegioInfas[,c("ID_t", "regio")])
#returns 0 if there are no duplicates
#If there are duplicates this command returns the index of the first duplicate

# existing regional levels are:
table(RegioInfas$regio)

# only keep housing level
RegioInfas = subset(RegioInfas, RegioInfas$regio ==  1)

# now you can enhance CohortProfile with regional data
CohortProfile <- read.neps("CohortProfile")
RegioInfas <- merge(CohortProfile, RegioInfas, by = c("ID_t","wave"), all = TRUE)
```

R 27: 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!
```

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 2 - RELEASE NOTES a.k.a CHANGE LOG
** changes and updates for release NEPS SC2 11.0.0
** (doi:10.5157/NEPS:SC2:11.0.0)
**
==================================================


==================================================
* Changes introduced to NEPS:SC2 by version 11.0.0 *
==================================================

General:
        - new data from survey wave 11 have been incorporated into the Scientific Use
            File

xPlausibleValues:
        - a new dataset has been added to the Scientific Use File, it provides ”
            Plausible Values” for the competence data
                stored in xTargetCompetencies


==================================================
* Changes introduced to NEPS:SC2 by version 10.0.0 *
==================================================

General:
        - new data from survey wave 10 have been incorporated into the Scientific-Use-
            File

CohortProfile:
        - the variables tx8600m/j, indicating the first date of answering the
            questionnaires, were renamed to tx8601m/j for
                reasons of consistency with the other NEPS starting cohorts

pTarget:
        - the variable ex20100, indicating the data lines recommended for linkage, were
            renamed to tx20100 for reasons of
                consistency with the other NEPS starting cohorts

xTargetCompetencies:
        - in contrast to the naming conventions for competence variables, the
            Scientific Use File includes the variables
                ’mag7q041_c’ and ’mag7q041_sc2g7_c’ with the latter referring to an
                    identical item that has been previously
                administered in SC3 and the former representing a new item that has
                    been administered in SC2 for the first
                time without any connection to the repeatedly administered item

xTargetCORONA:
        - this dataset was renamed from pTargetCORONA to xTargetCORONA and contains
            only variables from the additional
```

```
                CAWI survey in May 2020 on the Corona pandemic; the Corona-specific
                    information from the questionnaire
                of the regular survey wave 10 were integrated into the dataset pTarget


====================================================
* Changes introduced to NEPS:SC2 by version 9.0.0 *
====================================================

General:
        - all variables on dates of data collection (e.g., date of competency testing)
            were updated and stored together in the
                CohortProfile dataset (variables tx86***); the variables [intm] and [
                    inty] were removed from all other datasets

pParentCORONA:
        - new dataset with information by the parents from the additional online survey
             in May 2020 on issues related to the
                corona pandemic has been integrated in this SUF release
        - information on satisfaction with course of study, school or apprenticeship (
            variable p514010) 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

pTarget:
        - the online version of the questionnaire (CAWI) unfortunately contained an
            incorrect response scale for the variable
                t515051; accordingly, the values of this variable for the cases in
                    question were removed from the data set

xTargetCompetencies:
        - data on procedural metacognition have been released for wave 5 and updated
            for wave 6
        - WLE estimates and standard deviations have been added for waves 1, 3 and 5
            for the domain "vocabulary"


====================================================
* Changes introduced to NEPS:SC2 by version 8.0.1 *
====================================================

pParent:
        - variables for children's grades in German and Mathematics were missing in the
             last SUF release; these two variables
                are now added (p724101, p724102)


====================================================
* Changes introduced to NEPS:SC2 by version 8.0.0 *
====================================================

General:
        - new data from wave 8 have been incorporated into the Scientific Use File

EditionBackups:
        - this new dataset has been incorporated into the Scientific Use File for the
            first time; it contains raw values
```

```
                before data edition (for more details see the soon to be published Data
                     Manual update)


=====================================================
* Changes introduced to NEPS:SC2 by version 7.0.0 *
=====================================================

General:
        - new data from Wave 7 have been incorporated into the Scientific Use File
        - a new dataset "WeightsAsofGrade4" has been added (see below)

Weights:
        - The weighting dataset which contains the longitudinal weights for
            Kindergarten children only
          (WeightsKindergarten)is frozen and not updated any further due to small
              sample sizes.
        - An additional weighting dataset which starts in Wave 6 and is based on a
            composite design weight
          for all target children together again is provided (WeightsAsofGrade4).
        - Weights in "WeightsElementarySchool" and "WeightsAsofGrade4" are standardized
            , but not trimmed.
        - For more details see the accompanying weighting report for Wave 7
          "Samples, Weights and Nonresponse" by �Wrbach (2018);
          the report is available online in the documentation section of Starting
              Cohort 2 (www.neps-data.de).

=====================================================
* Changes introduced to NEPS:SC2 by version 6.0.1 *
=====================================================

Weights:
        - joint weights for students and parents have not been available in recent
            releases;
                they have been calculated by now and have been integrated into this
                    release

=====================================================
* Changes introduced to NEPS:SC2 by version 6.0.0 *
=====================================================

General:
        - translation for all meta data (variable and value labels, question texts, etc
            ) have been revised
        - meta data for all variables have been revised and updated where appropriate
        - additional wave 6 has been incorporated into the data
        - a new dataset "TargetMethods" has been added, reflecting methods information
            on the field procedure


=====================================================
* Changes introduced to NEPS:SC2 by version 5.1.0 *
=====================================================

        - in the xTargetCompetencies dataset, variable labels for variables "
            Uncorrected WLE-estimator mathematical competence"
                [mag1_sc1u] and "SE of uncorrected WLE-estimator mathematical
                    competence" [mag1_sc2u] accidentally had been swapped
                in earlier releases; this has been fixed;
                to implement the fix in earlier releases, these label issues can be
                    fixed by re-swapping labels again, for instance
                using the following piece of Stata syntax:
```

```
                * ---------- Begin Stata code ---------- *
                label language de
                label variable mag1_sc2u ‘”SE des unkorrigierten WLE-�Schtzers
                    Mathematischer Kompetenz”’
                label variable mag1_sc1u ‘”unkorrigierter WLE-�Schtzer Mathematische
                    Kompetenz”’
                char define mag1_sc1u[_lang_v_en] ‘”Uncorrected WLE-estimator
                    mathematical competence”’
                char define mag1_sc2u[_lang_v_en] ‘”SE of uncorrected WLE-estimator
                    mathematical competence”’
                * ---------- End Stata code ---------- *

        - two cases (ID_t: 300726 and 3019136) were mistakenly included in the
            xTargetCompetencies dataset of version 5.0.0, albeit the
                corresponding children did not participate in the tests (all variables
                    contained missing values for these observations).
                Thus, the respective data availability indicator in CohortProfile (
                    tx80522) should have been coded as 0, and the participation status
                indicator (tx80220) as 2, respectively;
                this has been fixed

        - in the CohortProfile dataset, the values of variable tx80501 (List of
            children / pupils: gender of child) accidentally had been completely
                swapped in version 5.0.0; this has been fixed.
                in version 5.0.0, these value issues can be fixed by re-swapping values
                     again, for instance using the following piece of Stata syntax:
                * ---------- Begin Stata code ---------- *
                recode tx80501 (1=2) (2=1)
                * ---------- End Stata code ---------- *

                The respective SPSS syntax is as follows:
                * ---------- Begin SPSS code ---------- *.
                RECODE tx80501 (1=2) (2=1).
                * ---------- End SPSS code ---------- *.

pCourseClass:
        - starting with this release, for the sake of consistency in data structures
            between panel cohorts, all information about school class
                contexts reported by the educators have been removed from the pEducator
                     dataset, and integrated to the new pCoureClass dataset instead;
                for end users to easily resolve multiple-educator-multiple-children-
                    scenarios, an indicator with a recommended observation for
                linkage [ex20100] is featured in the dataset as usual

pGroups:
        - the dataset ’Groups’ has been renamed to ’pGroups’

        - starting with this release, for the sake of consistency in data structures
            between panel cohorts, all information about Kindergarden
                groups reported by the educators have been removed from the pEducator
                    dataset, and integrated to the pGroups dataset instead;
                for end users to easily resolve multiple-educator-multiple-children-
                    scenarios, an indicator with a recommended observation for
                linkage [ex20100] is featured in the dataset as usual


pEducator:
        - as a result of the above, dataset ’pEducator’ starting with this release only
             features information on the educators (childminders / teachers)
                themself, and no longer on the Kindergarden group or class they teach;
                    for end users to easily resolve
                multiple-educator-multiple-children-scenarios, an indicator with a
                    recommended observation for linkage [ex20100] is featured
                in the dataset as usual

xTargetCompetencies:
        - linkage of WLE estimators for domains "maths" (mag*_sc1u), "reading" (reg*
            _sc1u), and "ict" (icg*_sc1u)
                had been errouneous in all previous releases; this has been fixed



==================================================
* Changes introduced to NEPS:SC2 by version 5.0.0 *
==================================================

CohortProfile:
        - the variable tx80501 (List of children / pupils: gender of child) suffered a
            coding error in wave 4. This has been fixed.

xTargetCompetencies:
        - in the current Scientific Use File the variable grg1_sc3 that represented the
             grammar sum score from wave 3 (grade 1) was
                deleted. Because the grammar test was aborted in some test groups due
                    to shortage of time, the sum score is not
                comparable between grous and thus not suitable for analyses. We
                    strongly recommend to use the wle score (variable grg1_sc1)
                instead.
        - the competency data for wave 1 to 4 suffered various minor coding errors.
            These were corrected in version 5.0.0.

        - competency scores and scales for domain "maths" surveyed in wave 3 (1st grade
            ) could not be delivered in due time for inclusion
                in version 3.0.0 and 4.0.0;
                starting from version 5.0.0, these items are completely integrated into
                     xTargetCompetencies


==================================================
* Changes introduced to NEPS:SC2 by version 4.0.0 *
==================================================

General:
        - translation for all meta data (variable and value labels, question texts, etc
            ) have been revised
        - meta data for all variables have been revised and updated where appropriate
        - additional wave 4 has been incorporated into the data

CohortProfile:
        - dataset CohortProfile erroneously contained the variable "ID teacher /
            educator" [ID_e]; however, it is not
                possible to directly assign teachers and educators to children in
                    Starting Cohort 2; assignment always has to
                take place using the pTarget dataset in order to corretly reflect the
                    educator-to-child relationship;
                thus, this variable has been removed

spSiblings:
        - in the spSiblings dataset, system missing values in variables "Sibling's date
             of birth - month" [p73221m] and
                "Sibling's date of birth - year" [p73221y] are incorrectly coded in
                    version 3.0.0, leading to implausible birth dates;
                this has been fixed
```

```
WeightsElementarySchool:
        - the english value label for value "3" of variable "Grouping variable for
          children in Kindergarten and school context" [group]
            is incorrect and should be "Target person from sample in wave 1,
                observation in school context";
            this has been fixed

WeightsKindergarten:
        - the english value label for value "3" of variable "Grouping variable for
          children in Kindergarten and school context" [group]
            is incorrect and should be "Target person from sample in wave 1,
                observation in school context";
            this has been fixed

pParent:
        - the concept of reflecting migrational background in NEPS SUFs has been
          improved in order to also represent migrants in 3.75th generation;
            thus, the older variables on migrational background [p400500_g1,
                p400500_g2,p400500_g3] in the pParent dataset have been renamed
                using
            the "v1" suffix [p400500_g1v1,p400500_g2v1,p400500_g3v1], and the new
                ones have been introduced

        - in version 3.0.0, there was an error in the derived values of target persons'
           CAMSIS scale (p731904_g15); this has been fixed


==================================================
* Changes introduced to NEPS:SC2 by version 3.0.0 *
==================================================

General:
        - starting with this release, all NEPS Scientific Use Files will ship with an
          additional, unicode-enabled Stata data set version;
            this version is only readable in Stata version 14 or younger, and is
                placed in the subdirectory "Stata14"
        - translation for all meta data (variable and value labels, question texts, etc
          ) have been revised and completed
        - meta data for all variables have been revised and updated where appropriate
        - additional wave 3 has been incorporated into the data, including observations
           from a sample refreshment in wave 3
        - regional information for German federal states have been added (for waves 1
          through 3, restrospectively) to the download SUF,
            more fine-grained information to the RemoteNEPS and onsite variant
        - weighting should be performed depending on the analyses focus; thus, the
          Weights data set has beend split up into two files:
            one for the analyses of waves one and/or two [WeightsKindergarten],
            the other for analyses of the full sample in wave 3 [
                WeightsElementarySchool]

CohortProfile:
        - in contrast to prior releases of Starting Cohort 2, a target person from now
          on is considered a "participant" [tx80220==1] only
            if he or she participated in competency testing; if there is only
                context information available on the target person, this
            person will be considered a "temporary drop out" instead [tx80220==2]

pInstitution:
        - in versions 2.0.0 and 1.0.0, variable h401820 has been treated as a single
          choice variable instead of a multiple choice variable set; this has been
          fixed,
                the (re-generated) muliple choice item variables are named h40182a,
                    h40182b, h40182c, h40182d

ParentMethods:
        - id variable ID_t was missing; this has been fixed

pTarget:
        - in wave 3 (grade 1) data, several educators could fill in distinct
            questionnaires about a single target person; this leads to two
            observations per target person
                for a total of 81 target persons; a recommendation marker [ex20100] has
                     beend added, recommending the observation with least item
                    nonresponse;
                thus, before merging data sets, unrecommended observations should be
                    dropped, or any other mechanism reducing these multiple lines to
                one single observation should be imposed
        - variable e67801c_g1 did contain value "0" instead of missing code -55 in case
             of missing values in one of the source variables; this has been fixed
        - variable eb10030 did contain an erroneous missing code "4" instead of "-98";
            this has been fixed

pParent:
        - variable p731702 was missing in version 2.0.0; this has been fixed
        - variable p743040 was missing in version 2.0.0; this has been fixed

spParentSchool:
        - for the sake of consistency between NEPS Starting Cohorts, the data set "
            spSchool" has been renamed to "spParentSchool"

xTargetCompetencies:
        - competency data for wave 2 stage-specific competency tests could not be
            delivered in due time by the responsible data editors for version 2.0.0;
                integration into the "xTargetCompetencies" data set now follows in
                    version 3.0.0, incorporating all competency assessment data up to
                    wave 3


===================================================
* Changes introduced to NEPS:SC2 by version 2.0.0 *
===================================================

General:
        - metadata for all datasets has been revised and updated where appropriate
        - variables now ship with a characteristic 'NEPS_instname' attached in Stata
            datasets, reporting the variable name used in the survey
        - wave 2 data has been fully integrated into the data
        - a new dataset "Weights" has been added, reflecting panel weights for the
            cohort; documentation is available online
        - a new dataset "spSchool" has been added, reflecting the target person's
            school history reported in the parent's interview
        - a new dataset "spSibling" has been added, reflecting the target person's
            siblings reported in the parent's interview
        - a new dataset "pTargetMicrom" has been added for onsite access, reflecting
            spatial data from "microm Micromarketing-Systeme und Consult GmbH"
        - a new dataset "pInstitutionMicrom" has been added for onsite access,
            reflecting spatial data from "microm Micromarketing-Systeme und Consult
            GmbH"
        - several bugfixes and enhancements have been integrated into this new release,
             influencing various variables;
                only the most important ones are listed in this change log

pParent:
        - as wave 2 data makes this a panel dataset, the filename has changed from ”
            xParent” to ”pParent”
        - the interview process in parent’s interviews does not guarantee unique ids
            for parents;
                thus, the identifier in this dataset is no longer ”ID_p”, but the
                    target person’s ”ID_t”
        - three variables with information about the target person’s migrational status
            have been calculated [p400500_g1, p400500_g2, p400500_g3];
                a working paper on the generation process and theoretical background is
                    forthcoming
        - values 5, 6 and 7 have been recoded to 7, 8 and 9 in ’Highest education
            qualification (ISCED)’ [p731802_g1]
        - values 5, 6 and 7 have been recoded to 7, 8 and 9 in ’Partner: Highest
            education qualification (ISCED)’ [p731852_g1]
        - value 96 has been recoded to -20 in ’Partner: (Highest) vocational education
            certificate’ [p731863] in accordance with official NEPS missing codes
        - ’Occupation (DKZ 2010)’ [p731904_g10] was erroneously in the dataset and has
            been removed
        - ’Partner: Occupation (DKZ 2010)’ [p731954_g10] was erroneously in the dataset
            and has been removed
        - ’Occupation (DKZ 1988)’ [p731904_g11] was erroneously in the dataset and has
            been removed
        - ’Partner: Occupation (DKZ 1988)’ [p731954_g11] was erroneously in the dataset
            and has been removed
        - EGP generation syntax was adjusted due to errors in the derivation syntax (
            particularly classes IVc and V) [p731904_g8, p731954_g8]
        - German EGP value labels have been corrected [p731904_g8, p731954_g8]
        - CASMIN [p731802_g2 & p731852_g2]: Class assignment slightly modified
        - ISCED [p731802_g1 & p731852_g1]: Civil servants of the medium grade are now
            identifiable
        - ’SDQ-Scale: Prosocial behaviour’ [p67801a_g1] has been corrected to only
            contain a sum score if all included items are non-missing

CohortProfile:
        - older weighting variables (’Standardized design weight’ [weight_design_std]
            and ’Design weight’ [weight_design]) are now
        deprecated and have been removed
        - the interview process in parent’s interviews does not guarantee unique ids
            for parents;
                as ”ID_p” therefore has been replaced by ”ID_t” and is no longer needed
                    to link datasets, it has been removed from CohortProfile
        - variable ’Test: survey day 1 (month)’ [test1m] has been renamed to [testm_w1]
        - variable ’Test: survey day 1 (year)’ [test1y] has been renamed to [testy_w1]
        - variable ’Test: survey day 2 (month)’ [test2m] has been renamed to [testm_w2]
        - variable ’Test: survey day 2 (year)’ [test2y] has been renamed to [testy_w2]

pEducator:
        - as wave 2 data makes this a panel dataset, the filename has changed from ”
            xEducator” to ”pEducator”
        - ’Further education, successful professional qualification (DKZ 2010)’ [
            e212821_g10] was erroneously in the dataset and has been removed
        - ’Further education, successful professional qualification (DKZ 1988)’ [
            e212821_g11] was erroneously in the dataset and has been removed

pInstitution:
        - as wave 2 data makes this a panel dataset, the filename has changed from ”
            xInstitution” to ”pInstitution”
        - ’Further education, successful professional qualification (DKZ 2010)’ [
            h212821_g10] was erroneously in the dataset and has been removed
        - ’Further education, successful professional qualification (DKZ 1988)’ [
            h212821_g11] was erroneously in the dataset and has been removed
```

```
pTarget:
        - as wave 2 data makes this a panel dataset, the filename has changed from ”
          xTarget” to ”pTarget”
        - ’SDQ-Scale: prosocial behaviour’ [e67801a_g1] has been corrected to only
          contain a sum score if all included items are non-missing
        - the scale of variable ’Helps other voluntarily’ [e67801i] has been
          erroneously reversed in generation of
              ’SDQ-Scale: prosocial behaviour’ [e67801a_g1]; this has been fixed

spChildCare:
        - the interview process in parent’s interviews does not guarantee unique ids
          for parents;
              thus, the identifier in this dataset is no longer ”ID_p”, but the
                  target person’s ”ID_t”
```