NEPS
National Educational Panel Study

Research Data

FDZ-LIfBi

Data Manual

NEPS Starting Cohort 8 – Grade 5 (2022)

Education for the World of Tomorrow

Scientific Use File Version 2.0.0

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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).

Contact

E-mail: fdz@lifbi.de
Web: https://www.lifbi.de/FDZ

Bibliographic Data

FDZ-LIfBi. (2026). *Data Manual of NEPS Starting Cohort 8 – Grade 5 (2022), Education for the World of Tomorrow, Scientific Use File Version 2.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

This data manual for the NEPS Starting Cohort 8 – Grade 5 (2022) "Education for the World of Tomorrow" has been prepared by the staff of the Research Data Center at the Leibniz Institute for Educational Trajectories (Forschungsdatenzentrum, FDZ-LIfBi). It represents a major collaborative effort. The contribution of the following persons is particularly acknowledged:

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

For their support in writing this manual, special thanks go to the following colleagues:

Elena Chincarini (LIfBi Bamberg)
Timo Gnambs (LIfBi Bamberg)
Katharina Molitor (TU Dortmund)
Lena Nusser (LIfBi Bamberg)
Ariane Würbach (LIfBi Bamberg)

Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3, 96047 Bamberg
Director: Prof. Dr. Cordula Artelt
Administrative Director: Dr. Stefan Echinger
Bamberg; August 15, 2026

Contents

1 Introduction

1.1 About this manual

This manual facilitates the work with data of the NEPS Starting Cohort 8 – Grade 5 (2022). It serves both as a first guide for getting started with the complex data and as a reference book. The primary focus is on aspects such as sampling and 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 from all previous survey waves as well as the new data from the most recently prepared survey wave – this manual is regularly updated and revised.

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 Research Data Center at the Leibniz Institute for Educational Trajectories (FDZ-LIfBi, Forschungsdatenzentrum). In the second chapter, the fundamental objectives of the NEPS Starting Cohort 8 (SC8) and its sampling strategy are briefly introduced. The main part of this chapter describes the sample realization of the first and second panel waves including field times, realized case numbers, survey modes, and the measurement of competence domains. The general principles of the data-editing process for the Scientific Use File as well as the applied conventions for naming 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 relevant data types, identifiers or ID variables, and portraits of all datasets in the Scientific Use File. These portraits include short syntax examples for merging variables of this dataset with variables from other datasets. The last chapter addresses some specific issues that should be considered when working with the data of NEPS Starting Cohort 8.

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 8, but they are transferable accordingly.

1.2 Data documentation

This manual cannot address all aspects of data documentation in detail. Therefore, numerous reports and additional materials containing background information (see Figure 1) are available for download on the website:

→ www.neps-data.de > Data and Documentation > Starting Cohort 8 > Documentation

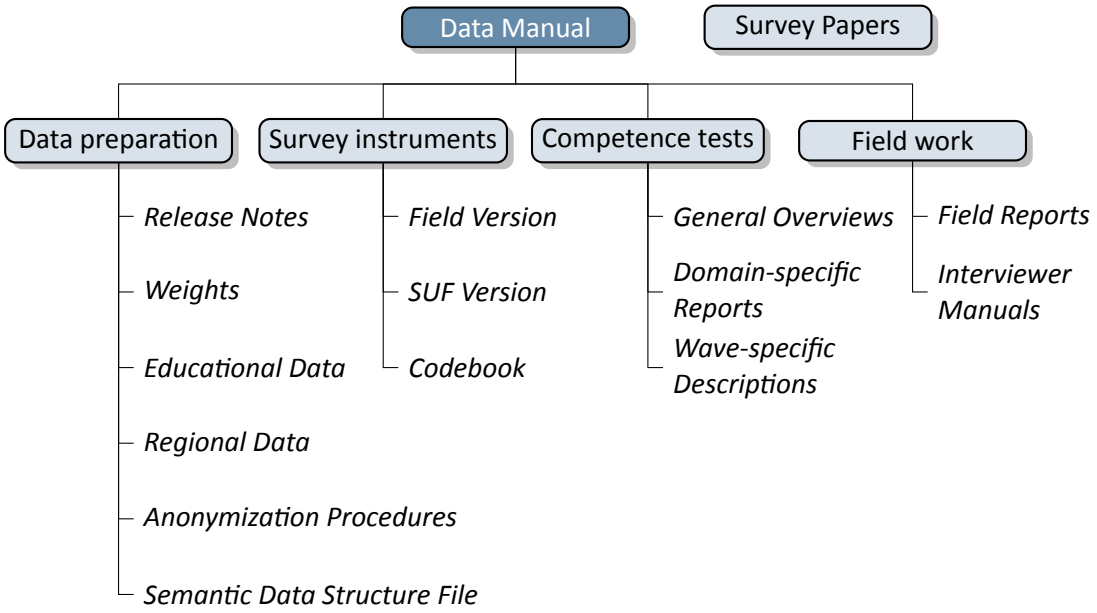

Figure 1: NEPS supplementary data documentation

Release Notes All Scientific Use Files are accompanied by release notes that list both relevant changes in the data compared to prior Scientific Use File versions and bugs eliminated or at least known of. For the latter, short syntax corrections are usually offered. Please consult these notes when working with the data. Section B.1 in the appendix shows the release notes current at the time the Scientific Use File was published; this information is constantly updated on the website mentioned earlier.

Weights The reports provide details regarding the sampling design, the weighting approach, and the generation of (wave-specific) weights.

Educational Data The report gives an overview of the generation of the derived educational variables ISCED-2011, CASMIN and Years of Education.

Regional Data The reports describe fine-grained regional indicators from commercial providers (microm, RegioInfas) that are available in the On-site environment – both in terms of the regional levels covered and the content with additional notes on how to merge it to the survey data (not available for Starting Cohort 8).

Anonymization Procedures The report describes the anonymization measures that were applied, including a list of all affected variables containing sensitive information, and gives an overview of the various ways to access the NEPS data.

Semantic Data Structure File The data package corresponds to the Scientific Use File but does not have any observations or filled data rows (*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. The Data Structure File can be downloaded from the website without any restrictions.

Survey instruments For each panel wave, the survey instruments are offered in the form of "Field versions" and "SUF versions". While the field versions consist of the originally deployed instruments (in German only), the SUF versions are enriched by additional information from the Scientific Use File such as respective dataset, the variable names and the value labels (in German and English). **Please note, that the competence test booklets are not publicly available**.

Codebook The codebook lists all variables and their corresponding labels (in German and English), together with their basic frequencies by wave. The structure is consistent with the data structure of the Scientific Use File.

Competence Tests Information about competence testing is provided in several reports, including general overviews and wave-specific descriptions. Usually, for each tested domain there is a brief description of the construct with item examples as well as a description of the data and the psychometric properties of the test.

Field Reports The field or method reports document the overall data-collection process conducted by the commissioned survey agencies, including details about survey preparation, interviewer training, respondent tracking, initial contacts, incentives, and sample realization (in German only).

Interviewer Manuals These wave-specific manuals are a collection of instructions for the interviewers. They exemplify the interview process and the content of the questionnaire modules (in German only; not available 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 can be searched for and downloaded from the LIfBi website:

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

Additional documentation material might be available at the documentation website mentioned at the beginning of this chapter.

1.3 Data releases

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

File Format

The Scientific Use Files are offered in Stata and SPSS format with variable and value labels in German and English. In the SPSS format, separate data files are available for each language. Data stored in the Stata format contain both languages within one file; the switch is induced by the Stata command:

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

Versioning and Digital Object Identifier

With every new release, the existing data files of a Scientific Use File are either extended – usually by information from a new survey wave – or updated with changes due to larger or smaller corrections. The three-digit version number indicates the number of panel waves of available data, the frequency of major updates, and the frequency of minor updates. The version number is part of the name of the Scientific Use File, its data files (see Table 3), and the associated Digitial Object Identifier (see Table 1).

All NEPS Scientific Use File releases are registered at `DataCite` and equipped with a unique *Digital Object Identifier* (DOI; see Wenzig, 2012). This DOI has two main functions: It enables researchers to cite the used NEPS data in an easy and precise way (see Section 1.5), which is considered good scientific practice and essential for any replication analysis. The DOI also leads to a landing page with further information about the Scientific Use File and the data access options.

Table 1: Release history of Scientific Use Files in Starting Cohort 8

SUF Version	DOI	Date of release
2.0.0 (current)	`doi:10.5157/NEPS:SC8:2.0.0`	**2026-07-23**
1.0.0	`doi:10.5157/NEPS:SC8:1.0.0`	2025-05-07

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. It is granted upon the conclusion of a *Data Use Agreement* with the FDZ-LIfBi. The existence of a valid agreement entitles the parties involved to work with all NEPS Scientific Use Files, i. e., the full data portfolio with seven starting cohorts and NEPScomp (see Section 1.8) is available to them.

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 have signed the agreement. Submit the completed document by e-mail or mail. Further instructions and the relevant forms are provided on the website:

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

- After approval by the FDZ-LIfBi, each registered person receives an individual user name and a password in separate e-mails to log in to the website for downloading the NEPS data:

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

- If access to more sensitive data via RemoteNEPS or On-site data use (see below) is required, a corresponding Supplemental Agreement is necessary.

- Another form is offered to state changes to an existing Data Use Agreement, such as the later addition of further project participants or an extension of the project duration.

Modes of data access

Three modes of accessing the NEPS Scientific Use Files are possible to support the full range of researchers' interests regarding data utility while complying with national and international standards of confidentiality protection. Each mode corresponds to a Scientific Use File version that is different in terms of the availability of sensitive information (anonymization) and the level of control over data use (monitoring).

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

While working with RemoteNEPS requires biometric authentication and Internet access, the On-site use of NEPS data requires a guest stay at the data security room in Bamberg and in-person check-in with the FDZ-LIfBi staff. The monitoring of data work in the secure Remote and On-site environments mainly involves the control of all output files prior to their transfer ("export") to the data user. More details about the different access modes are given at:

→ www.neps-data.de > Data Access

Sensitive information

The Download version of a Scientific Use File contains the least amount of information. For instance, institutional context data (`pInstitution`, see Figure 8) or variables to distinguish between the federal states (*Bundeslandkennung*, see Section 1.7) are only available in RemoteNEPS and On-site. Certain variables containing sensitive information that might increase the risk of re-identification 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 most restrictive access version of On-site data use, e. g., fine-grained regional indicators or open text entries. For a more specific overview:

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

This concept of *nested data dissemination* translates into an onion-shaped model of data provision. The most sensitive On-site level represents the outer layer with the Remote and Download levels being subsets. That is, any information contained within a less sensitive access level is also included in the corresponding higher level(s), but not vice versa. The report on "Anonymization Procedures" (see Section 1.2) illustrates this model and outlines the related logic of NEPS variable naming (see Section 3.2.1), including a list of all variables concerned in the Scientific Use File.

1.5 Citation guidelines

Referencing the use of NEPS data is essential for a good scientific practice and for revealing the scientific value of the National Educational Panel Study. The following citation guidelines apply to all publications based on analyses of data from the NEPS Starting Cohort 8. First of all, it is obligatory to specify the **Scientific Use File** utilized with its DOI as follows:

> NEPS Network. (2026). *National Educational Panel Study, Scientific Use File of Starting Cohort 8 – Grade 5 (2022)*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC8:2.0.0

Second, the **NEPS study** should be acknowledged in the publication:

> 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.

Third, the **reference article** is to be listed in the bibliography:

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

Authors of any kind of publications based on NEPS data are requested to inform the FDZ-LIfBi by sending an e-mail with the bibliographic details to `fdz@lifbi.de`. All known publications are listed on the respective website:

→ `www.neps-data.de` > `Publications`

For references to **documentation materials** (e. g., this manual), the following citation templates are recommended:

> FDZ-LIfBi. (2026). *Data Manual of NEPS Starting Cohort 8 – Grade 5 (2022), Education for the World of Tomorrow, Scientific Use File Version 2.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

If no author is given (e. g., survey instruments), a universal "NEPS Network" should be taken instead:

> NEPS Network. (2026). *Starting Cohort 8 – Grade 5 (2022), Wave 2, Questionnaires (SUF Version 2.0.0)*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

For "externally" produced documents (e. g., field reports), standard citation guidelines apply:

> Hellrung, M., Hillen, P., Hugk, N., Meyer-Everdt, M., Sievers, U., & Tusch, S. (2025). *Feld- und Methodenbericht der IEA Hamburg zur NEPS-Teilstudie A104*. Hamburg, Germany: IEA.

1.6 General rules and recommendations

Working with NEPS data is bound to a couple of general rules. Data users have to confirm these rules through their signature on the Data Use Agreement. The just mentioned obligation to cite the Scientific Use File utilized for analysis and to report any kind of publication resulting from the use of NEPS data are two examples. The major part of rules, however, 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 accidental re-identification, the FDZ-LIfBi (see Section 1.9) has to be informed immediately and all individual data gained therefrom have to be kept secret.
- *Avoidance of data disclosure:* The NEPS data portfolio is 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 applicant and further project members that are mentioned by name). Any transfer of the data to third parties or use for commercial and other economic purposes is not permitted. NEPS data must be handled with strict confidentiality!

- *Regulations on using the federal state label:* For NEPS data collected in the context of schools or higher education institutions, it is forbidden to use federal State-related information, whether directly or indirectly contained in the data, for analyses that aim to (a) make direct comparisons between German federal states (*Bundesländer*), (b) draw specific conclusions about an individual federal state, and (c) to reconstruct the federal state affiliation of individuals, households, or institutions. Any kind of ranking between the federal states based on NEPS data is not permitted (see Section 1.7).

It is important to note that a violation of these rules may lead to severe penalties as stated in the Data Use Agreement. If there is any doubt or question regarding the given regulations, please contact the FDZ-LIfBi (see Section 1.9). The same applies in case of encountering any issues concerning data quality or security leaks with regard to data protection.

Recommendations

- *As a matter of course:* Always be critical when working with empirical data. Although great efforts are made to ensure the integrity of the research data, absolute correctness cannot be guaranteed. The FDZ-LIfBi welcomes information about any identified problems or errors.
- *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 was collected and processed. All materials are available online (see Section 1.2).
- *Facilitated handling of the data:* Use the assistance tools provided. Several user services are available to support NEPS data analyses – from specific Stata commands (e. g., for an easy recoding of missing values) to a variable search platform (e. g., for an interactive exploration of all instruments) and an online discussion forum (e. g., for asking specific questions). These tools are also offered at the LIfBi web portal (see Section 1.8).

1.7 On using the federal state variable *(Bundeslandkennung)*

Using the federal state variable or label in the NEPS data collected in connection with schools or higher education institutions is permitted only if it is exclusively for:

- control purposes in order to incorporate it as a covariate; 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 variable in the starting cohorts of schools and higher education institutions is provided only via Remote access (*RemoteNEPS*) and via guest working stations at the LIfBi in Bamberg (*On-site*). The respective analysis results are reviewed by staff of the FDZ-LIfBi before being passed on to the researcher in a password-protected environment (see Section 1.4).

1.8 User services

Along with the comprehensive data documentation, there are several user services to support researchers working with the NEPS data. First and foremost, the FDZ-LIfBi maintains a regularly updated and enhanced website with detailed information on all Scientific Use Files, a complete list of registered NEPS data analysis projects, notices about NEPS-related events, and a newsletter. All subsequently introduced services are free of charge and online available.

→ www.neps-data.de > NEPS

NEPScomp

NEPScomp is designed particularly for the use in university courses and in student theses. It features a number of simplifications with regard to data structure and data handling compared to the regular Scientific Use Files. The “comp” in the title – short for “compenedium” – refers to the explicit claim to offer a compact and easy-to-understand NEPS data package that still enables empirically valid findings on a wide range of questions in the field of educational research. NEPScomp is part of the entire NEPS data portfolio. This means that, given a valid Data Use Agreement (see Section 1.4) exists, the data can be downloaded together with the regular Scientific Use File for the individual starting cohorts.

→ www.neps-data.de > NEPScomp

Variable Search

The *Variable Search* facilitates an interactive and quick full text search through all instruments of released NEPS surveys, including survey and 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. It has some helpful extra features such as displaying the presence of the selected variable in all NEPS starting cohorts, the answering scheme, relevant references or sources, and more. As a web application, the service relies on the most up-to-date information; any correction in the metadata is thus instantly visible.

→ www.neps-data.de > Variable Search

Online Forum

The *Forum4MICA – Making Information Commonly Available* is an open discussion platform for data users as well as for persons who are just searching for information. The forum is joined by various Research Data Centers with their data collections, among them the FDZ-LIfBi with the NEPS and other research data resources. It provides the opportunity to directly exchange with NEPS staff members and with other researchers in a transparent dialogue. It is recommended that users first search through the content first when struggling with NEPS issues or whenever help is needed with specific data-related matters. If no solution can be found there, the question should be posted in the forum. Active participation is highly encouraged and requires no more than a one-time registration with a username and an e-mail address. The NEPS research community (and beyond) benefits from a broad participation.

→ https://forum.lifbi.de

Data Trainings

The FDZ-LIfBi conducts a series of regular NEPS data trainings in the form of online courses. 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 start working 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. The schedules of all current training courses are provided on the website. There is also information about registering for a course, which can be taught in German or English – depending on the participants' needs.

→ www.neps-data.de > Data Trainings

NEPS Tutorials

The *NEPS Tutorials* are short video presentations that serve as visual support and supplement to the NEPS data trainings and the documentation materials. They focus on general information about the NEPS and its cohorts or the data structure of the Scientific Use Files, but also on specific topics such as the combination of information from different datasets ("merging") or the survey data linked to administrative data ("NEPS-ADIAB"). The tutorials are currently available in German only.

→ www.neps-data.de > NEPS Tutorials

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 (PV) is suitable for more precise inferential statistical tests in correlation and mean value analyses. The R package *NEPSscaling* enables data users to generate their own PV 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.

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

NEPStools

The *NEPStools* is a collection of specific Stata commands. 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 (see Section 3.3). Another example is the **infoquery** command that displays extended characteristics of the variable(s) such as the question text and the initial variable name in the field instrument. NEPStools can be installed from the FDZ-LIfBi 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.

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

1.9 Contact

The Research Data Center at the Leibniz Institute for Educational Trajectories (FDZ-LIfBi) accounts for large parts of the NEPS data preparation and documentation, for the data dissemination, and for the user support. We appreciate any feedback to further improve the research data and services offered.

Please contact the FDZ-LIfBi with questions, comments, requests, and suggestions.

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

2 Sampling and Survey Overview

2.1 Education for the world of tomorrow

The lower secondary level plays a connecting role between elementary school and the general or vocational upper secondary level. This educational stage is associated with many processes that raise particular questions about:

- reasons for choosing a certain type of school or school track (*Schulzweig*)
- transitions to other school types
- conditions for successful learning or competence acquisition
- consequences of grade retentions

With *Education for the World of Tomorrow*, the NEPS has launched a second cohort in lower secondary level from grade 5 onwards, twelve years after Starting Cohort 3. This cohort succession allows for investigating the effects of social and educational policy changes over the last decade on everyday school life. The new NEPS Starting Cohort 8 focuses on topics such as:

- use of digital media in education and in students' everyday lives
- digital competencies and role of social media
- relevance of political participation and involvement in civic engagement
- changes in the school system and integration of increasingly diverse groups of students

The general scope of this cohort still covers the students' development in various competence domains as well as the educational processes, educational decisions and returns to education over the course of the secondary school level – each in combination with a range of contextual aspects (e. g., institutional and family factors). On the one hand, this opens up a highly interesting comparative perspective on Starting Cohort 3. On the other hand, the Starting Cohort 8 offers an empirical basis for completely new research approaches. The survey design enables, for example, a more specific view on the school as learning context (by including the entire teaching staff in the survey) and on the level of reading and math competencies (by linking the NEPS measures to the IQB Bildungstrends based on identical tests and schools).

The aim of this panel study is to establish and provide research data for a better understanding of the challenges and opportunities of education for the world of tomorrow. For this purpose, the NEPS Starting Cohort 8 follows a sample of representatively selected fifth-graders who attended regular and special schools in 2022/23 and were willing to take part in annual surveys and competence tests. Parallel to the target respondents, surveys are conducted with their parents, teachers and school principals or members of the school administration team.

2.2 Sampling strategy

The target population of Starting Cohort 8 includes students of the fifth grade in German schools offering lower secondary education in the school year 2022/2023.[1] Access to the target respondents was organized through the schools in which the students were enrolled. The initial sample has been drawn using a stratified procedure in which schools were selected at random. Within the selected schools, all fifth-grade classes were invited to participate in the NEPS study, unless the schools made their own selection of classes.

In the first stage, the selection was based on a list of all regular schools in Germany as **primary sampling units** (PSU). This school frame contained only schools of the general education system (e. g., no vocational schools). It was compiled using up-to-date school registers for the school year 2020/2021 from the Statistical Offices of all 16 German federal states. In order to adequately reflect the diversity of the federal state-specific school systems, all schools in the frame were classified according to their type or track.[2] The school type or track – *Gymnasium, Realschule, Hauptschule, Schule mit mehreren Bildungsgängen, Gesamtschule, Schulformunabhängige Orientierungsstufe, Förderschule* – served as an explicit stratification criterion. Other characteristics were considered for implicit stratification:

- federal state
- degree of urbanization or regional classification
- sponsorship (public vs private)

From the total list of 11,402 schools or school tracks (excluding special schools with a focus other than learning and schools with a non-German language of origin in Hesse), 575 original schools were selected for grade 5, of which 450 were regular schools and 125 were special schools with a focus on special educational needs in the area of learning. For each selected school, up to nine replacement schools with identical characteristics in terms of federal state, sponsorship and degree of urbanization as well as similar class sizes were drawn from the list. In sum, the gross sample included 5,068 schools, thereof 4,478 regular and 590 special schools.

Participation in the NEPS is always voluntary – that applies to all participants including the schools. Despite the support of the Ministries of Education in several federal states, the recruitment of a sufficient number of schools proved to be a particular challenge. If a school of the original sample refused to take part in the study, the loss was attempted to be compensated for by one of the structurally equivalent replacement schools. Even with this replacement, only 270 schools (thereof 29 special schools) finally agreed to take part, corresponding to a response

1 In Berlin and Brandenburg, elementary schools were also recruited and included in the survey due to the specific transition regulations in the school system there.

2 An individual school could be assigned to a certain type of school only; however, it could also offer different tracks (*Schulzweige*). The sampling was not based on institutions, but on school tracks. Accordingly, schools with several tracks were represented several times in the sampling frame.

rate of 5 percent at the institutional level and 47 percent of the originally planned school sample.[3]

In the second stage of the sampling process, the survey agency asked all recruited schools for information on their classes in grade 5, including the number of students per class. Every fifth-grade class was eligible to participate. If a school made a selection from among its fifth-grade classes, only these classes were considered. In the participating classes, all students were asked to become part of the study. Thus, the students constitute the **secondary sampling units** (SSU).[4]

The target group for Starting Cohort 8 consists of **students**, starting with the first survey wave in fifth grade. In order to actually be allowed to participate, one legal guardian of the child had to give written consent. Only students for whom a fully completed consent form was available on the day of the first survey were eligible to take part. The individual forms were distributed to the students via the schools and collected again there. Out of the gross sample of 20,535 registered students in the selected fifth-grade classes, a total of 6,141 students (thereof 241 from special schools) met the requirements for being included in the NEPS Starting Cohort 8.

- For the first panel wave, data is available for 5,763 out of these 6,141 students; that is, they had taken at least one competence test and/or completed the student questionnaire. This corresponds to a participation rate of 93.8 percent (see also Section 2.4.1).
- For the second panel wave, the Scientific Use File contains data from 5,215 students who completed at least one part of the survey. Based on a sample size of 6,011 individuals, this results in a response rate of 86.8 percent (see also Section 2.4.2).

Students who could not be reached at the school on the regular survey day or, if applicable, an alternate date (e. g., due to illness), or who could no longer be surveyed at the originally selected "NEPS school" (e. g., due to a change of school or the school's refusal to continue participating in the study), were individually followed or tracked and asked for information through alternative channels – such as online questionnaires, telephone interviews, and/or paper questionnaires (see Section 5.3 and the respective field reports for details).

3 In order to be able to compare the competencies of the students participating in the NEPS at a later date (end of lower secondary level) with the competence data collected from grade 9 students as part of the IQB Bildungstrends study (IQB-BT, see https://www.iqb.hu-berlin.de/bt), the school sample also included schools that had participated in the IQB-BT with grade 9 in spring 2022. The gross sample contained 1,080 schools across the different tracks that had participated in the IQB Bildungstrends study with grade 9 in spring 2022. Of these 1,080 schools, 83 agreed to be part of the NEPS study.

4 Further information on the initial recruitment process and the sampling procedure for Starting Cohort 8 can be found in the PAPI/CASI field report on the first survey wave among students in schools (in German only, see Hellrung et al., 2025). The implications of the sampling design for the derivation of weighting variables are described in the report by Konrad et al., 2025. Both papers are part of the data documentation and are available online (see Section 1.2).

Context information

The **school context** – and thus the students' institutional learning environment – is taken into account through separate surveys of teachers and school administrations. The schools receive cover letters, which are forwarded to the relevant teachers and include all necessary information for the individual access to the online instrument. Participation in the study is, of course, also voluntary for all school context respondents. The survey consists of various role-specific question modules. If a teacher holds multiple roles for the participating "NEPS classes" at a school – for example, German teacher and class teacher, or math teacher and school administrator – the corresponding modules are presented to this person in sequence. Since the questions for teachers, apart from a general questionnaire section, always refer to specific classes, it may happen that a person is asked to complete the German teacher module multiple times, for example, if that person is responsible for teaching German in several "NEPS classes" at that school. Conversely, it is also possible that information from multiple teachers is available for a single German "NEPS class", for example, if co-teaching is implemented. School administrators or members of the school administration team answered a module with questions on school-related topics.[5]

- In the first survey wave, teachers of German and mathematics, as well as class teachers of the "NEPS classes", were surveyed with the corresponding function-specific modules. In addition, the entire teaching staff and the school administration of all participating schools were invited to participate in the survey (see also Section 5.4 and Section 2.4.1).[6]
- For participation in the second survey wave, only classroom teachers were asked to participate, along with school administrations that had not already taken part in the previous survey wave. In addition to the standard teacher questionnaire modules, the classroom teachers were also asked to provide individual assessments of their "NEPS students" (see also Section 5.5 and Section 2.4.2).

Information about the **family and home context** of the students in Starting Cohort 8 is primarily collected through repeated interviews with their legal guardians. These interviews primarily focus on education-related aspects and the student's school career, but also cover parental support for the child, the child's health, satisfaction with school, language use in the family, household resources, and various sociodemographic characteristics. Initially, the biological or social parent of a target child was invited to participate in the study who has the greatest familiarity with the child's school-related matters. In the vast majority of cases, this was the biological mother (approximately 80 percent in the first survey wave). Participation in the parent interviews is also voluntary. Consent for the parent's participation was obtained separately from consent for the child's participation. A change in the surveyed parent was possible at a later date, provided that appropriate consent was given and the person was both a legal guardian and a biological or social parent of the target child (see also Section 5.6).

5 Due to the increased risk of re-identification, the respective dataset (`xInstitution`) is only available in the RemoteNEPS and the On-site version of the Scientific Use File.

6 Inviting the entire teaching staff to participate in the survey represents a significant innovation in the study design compared to the NEPS Starting Cohort 3 – Grade 5 (2010).

- The parent interviews in the first survey wave were done as computer-assisted telephone interviews with language options in German, Russian, and Turkish (see also Section 2.4.1).
- In the second survey wave, the parent survey was administered via a web-based online questionnaire, which could be completed in German, Russian, and Turkish again (see also Section 2.4.2).

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. Measurements are carried out across different waves in all NEPS starting cohorts covering **domain-general competencies** and **domain-specific competencies** as well as **metacompetencies** and – for some cohorts – also **stage-specific competencies**.

Data from the competence tests pass through an editing process before they are 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 are available in several reports on the data documentation website (see Section 1.2).

- The individual and generated scores for students at *regular schools* are compiled in the dataset `xTargetCompetencies`.
- The values for the tests that were administered at *special schools* in the Starting Cohorts 3, 4, and 8 are provided in the dataset `xTargetSpecialNeedsCompetencies`. The competence assessments in special schools are generally based on shorter sets of tasks specifically tailored to the needs of the students there. The data documentation usually provides reports with detailed information on how the competence estimates derived from these tests can be interpreted, as well as on the comparability of competence measures between regular and special education schools (see Section 1.2).
- The Scientific Use File contains another competence dataset called `xPlausibleValues`, which offers exemplary variables with "plausible values" that were generated using the freely available R package *NEPSscaling* (Scharl and Zink, 2022; see also Section 1.8).

All competence data are 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 competence domain, the target group, the testing modus, and the kind of scoring – they also indicate the repeated administration of a test item in a different wave or starting cohort (see Section 3.2.2 for details).

Table 2 on the following page shows the schedule of competence measures in NEPS Starting Cohort 8 with domains by waves and test modes. If the information goes beyond the survey

waves currently available in the Scientific Use File, it reflects the latest known state of planning. A few wave-specific notes should be taken into account in addition to the table:

- Wave 1: The test used in wave 1 to measure reading competence (re) in special schools can be used for comparative analyses with data from regular schools in Starting Cohort 8 (see Gnambs, 2025a and Gnambs, 2025b).
- Waves 1 and 4: The tests for assessing verbal and non-verbal reasoning (vi, ni) were administered only in special schools and not in regular schools (see Gnambs, 2025c). The same tests were already applied in Starting Cohort 3 among students at special schools and can therefore be used for cross-cohort analyses.
- Waves 1 and 5: There is a randomized allocation of IQB Bildungstrend tests on reading and mathematics competence to a split sample (50% rb + 50% mb). In wave 5, the same domain will be administered as in wave 1.
- Wave 2: The L1-Test for Russian and Turkish language is applied to target persons of a corresponding migration background only.

Table 2: Schedule of competence tests

		2022/23 **Wave 1** Grade 5	2023/24 **Wave 2** Grade 6	2024/25 **Wave 3** Grade 7	2025/26 **Wave 4** Grade 8	2026/27 **Wave 5** Grade 9
Domain-General Competencies						
DGCF: Cognitive Basic Skills	dg	—	—	C **C**	—	—
General Knowledge*	gk	—	C **C**	—	—	—
Verbal Reasoning	vi	**P**	—	—	**C**	—
Nonverbal Reasoning	ni	**P**	—	—	**C**	—
Domain-Specific Competencies						
Reading Competence	re	P **P**	—	C **C**	—	C **C**
Reading Competence (IQB-BT)	rb	P	—	—	—	P
Reading Speed	rs	—	—	C **C**	—	—
Vocabulary: LC at Word Level*	vo	—	C **C**	—	—	—
Mathematical Competence	ma	P	—	C	—	C
Mathematical Competence (IQB-BT)	mb	P	—	—	—	P
Scientific Competence*	sc	—	—	—	C	—
Native Language Russian/Turkish: LC*	nr/nt	—	C	—	—	—
Metacompetencies						
ICT Literacy*	ic	—	C	—	C	—
Digital Competence*	dc	—	C **C**	—	C **C**	—
Civic Literacy*	cl	—	—	C	—	—

Notes: P = Paper-Based Test (proctored), C = Computer-Based Test (proctored). The non-colored boxes with capital letters refer to tests conducted with students in regular schools; the blue highlighted boxes indicate tests administered to students in special schools.
* Following several competence tests (gk, vo, sc, nr/nt, ic, dc, cl), the target persons were also asked to evaluate their own test performance ("Procedural Metacognition", mp).

2.4 Survey overview and sample development

This section informs about the progress of the NEPS Starting Cohort 8 sample. For each panel wave of the current Scientific Use File, a short characterization with information on field times, groups of respondents, number of realized cases, survey modes, and the survey institute(s) responsible for collecting the data is given. The information is essentially an excerpt from the respective field reports, which provide much more detailed insights into the implementation of the data collection procedures for each panel wave (in German only; see Section 1.2). The case numbers in the figures are taken from the Scientific Use File; minor discrepancies with the numbers in the field reports cannot be ruled out due to the comprehensive preparation of the "raw" data material.

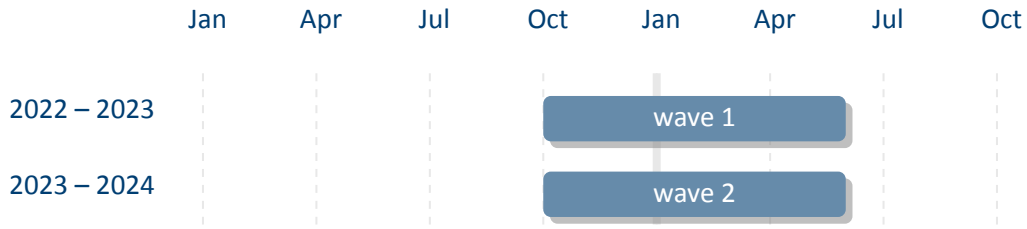

Figure 2: Panel progress of Starting Cohort 8

2.4.1 Wave 1: 2022/2023

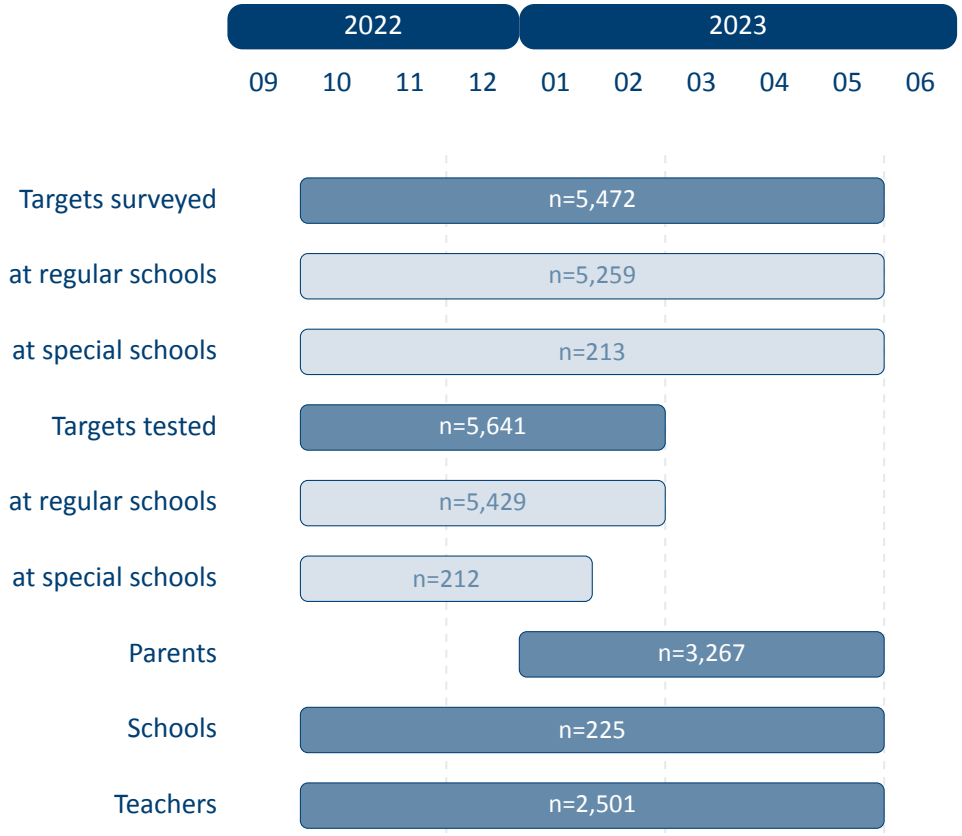

Figure 3: Field times and realized case numbers in wave 1

- **Target persons** Fifth-graders at panel start 2022/23
  - *Grade 5 students at regular and special education schools*

    Sample Stratified cluster sampling (see Section 2.2):
    Size-proportional random selection of regular schools at lower secondary level (and also at elementary level in Berlin and Brandenburg) with stratification regarding school track, federal state, degree of urbanization, and sponsorship
    Full survey of 5th grade classes in schools; if requested, a selection of classes in 5th grade was made by the school
    All students of the participating classes were invited to the survey

    Modus Computer-assisted self-interviewing questionnaires (CASI) + Written competence tests (PAPI) completed in class context = school context

Computer-assisted web interviewing questionnaire (CAWI) for students who could not be reached at school on the day(s) of the survey = individual field

Testing regular schools: Reading + Reading (IQB-BT), Mathematics + Mathematics (IQB-BT)
special education schools: Verbal Reasoning, Nonverbal Reasoning, Reading

Individual survey 96 students could not be reached at the school for the survey, but answered the questionnaire online via CAWI (n=95 from regular schools, n=1 from special education school)

- **Context persons**
  - *Parents*

    Sample One legal guardian per target child = biological or social parent

    Modus Computer-assisted telephone interviews (CATI)

  - *Teachers*

    Sample Teaching staff of all schools with participating classes + class teachers of the target children and their teachers for German and mathematics

    Modus Computer-assisted web interviewing questionnaire (CAWI)

    Note Teacher appear several times in the `Course#` datasets if they have indicated responsibility as a class, German or math teacher for more than one "NEPS class" and have answered the corresponding modules accordingly

  - *School principals/administrations*

    Sample Principals or school administration of all schools with participating classes

    Modus Computer-assisted web interviewing questionnaire (CAWI)

    Note 225 is the number of schools for which responses are available from the principal or school administration
    for some schools, several members of the school administration have answered the corresponding questionnaire module

- **Data collection**
  - *Responsible for conducting the survey*

    CASI/PAPI students at school IEA DPC – Data Processing and Research Center, Hamburg

    CAWI school staff + students not reached at school LIfBi – Surveytechnology, Bamberg

    CATI parents infas – Institute for Applied Social Sciences, Bonn

2.4.2 Wave 2: 2023/2024

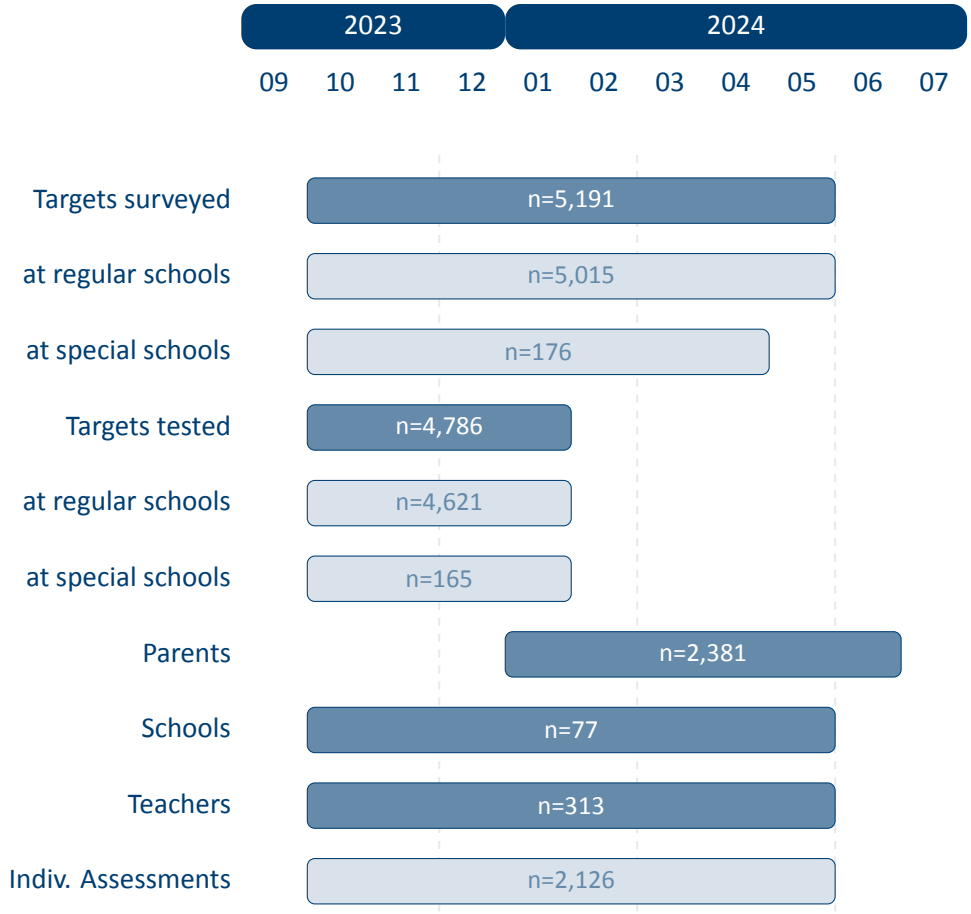

Figure 4: Field times and realized case numbers in wave 2

- **Target persons** Fifth-graders at panel start 2022/23
  - *Grade 6 students at regular and special education schools*

    Sample Stratified cluster sampling (see Section 2.2):
    Same sample as in wave 1 – unless participation has been withdrawn in the meantime or there has been another final dropout

    Modus Computer-assisted self-interviewing questionnaires (CASI) + Technology-based Testing (TBT) in class context = school context
    Computer-assisted web interviewing questionnaire (CAWI) for students who could not be reached at school on the day(s) of the survey = individual field

Computer-assisted telephone interviews (CATI) + Computer-assisted web interviewing questionnaire (CAWI) or alternatively written questionnaire (PAPI) for students who could no longer be reached at the "NEPS school" = individual tracking

Testing regular schools: ICT Literacy, Digital Competence, Vocabulary (Listening Comprehension at Word Level), General Knowledge or Russian/Turkish (Listening Comprehension, only students with corresponding language background)
special education schools: Digital Competence, Vocabulary (Listening Comprehension at Word Level), General Knowledge

Individual survey 214 students could not be reached at the school for the survey, but answered the questionnaire online via CAWI (n=210 from regular schools, n=4 from special education schools)
another 215 students were surveyed in the course of the individual tracking (n=209 from regular schools, n=6 from special education schools)

- **Context persons**
  - *Parents*

    Sample One legal guardian per target child = biological or social parent
    Preference for surveying the same parent as in Wave 1, but option to change the respondent – if biological or social parent status, primary responsibility for the target child's daily and school-related needs, and shared household context with the target child

    Modus Computer-assisted web interviewing questionnaire (CAWI)

  - *Teachers*

    Sample Class teachers of the target children or participating "NEPS classes" only

    Modus Computer-assisted web interviewing questionnaire (CAWI)

    Note Survey included individual assessments of all "NEPS students" in the class
    275 class teachers provided individual assessments for 1,822 "NEPS students"
    Some students were assessed more than once by different class teachers, so the total number of completed assessment questionnaires is n=2,126

  - *School principals/administrations*

    Sample Only principals or school administrations of schools for which no completed questionnaire was available from the first wave

    Modus Computer-assisted web interviewing questionnaire (CAWI)

- **Data collection**
  - *Responsible for conducting the survey*

    CASI/TBT students at school IEA DPC – Data Processing and Research Center, Hamburg

CAWI school staff + students not reached at school + individual tracking + parents LIfBi – Surveytechnology, Bamberg

CATI/PAPI individual tracking 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 FDZ-LIfBi). 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 source 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). This strategy is based on the assumption that it is 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 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 8, although some of the examples refer to other NEPS starting cohorts.

3.1 File names

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

Table 3: Naming conventions for NEPS data files

Element	Definition
SC[1–8]	**Indicator for the starting cohort and launch year of the panel** 1 = Newborns (2012) 2 = Kindergarten (2011) 3 = Fifth-grade students (2010) 4 = Ninth-grade students (2010) 5 = First-year university students (2010) 6 = Adults (2007) 8 = Fifth-grade students (2022)
[filename]	**Meaning of the file name** *Prefix*: x = cross-sectional file; sp = spell file; p = panel file *Keyword*: indicates the content of the file (e. g., pTarget contains panel data with regard 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 (syntax updating not necessary)

For instance, SC8_CohortProfile_D_2.0.0 refers to the generated *CohortProfile* dataset of *Starting Cohort 8* in its *Download* version of the Scientific Use File release *2.0.0*.

3.2 Variables

The naming conventions for variables in a NEPS Scientific Use File aim to ensure maximum consistency both between the panel waves and between the different starting cohorts. The names refer to various characteristics and thus provide the data user with a first orientation regarding the contents of the variables. The principles of these naming conventions are exemplified in Figure 5. It has to be noted that separate conventions apply for survey variables and for competence variables.

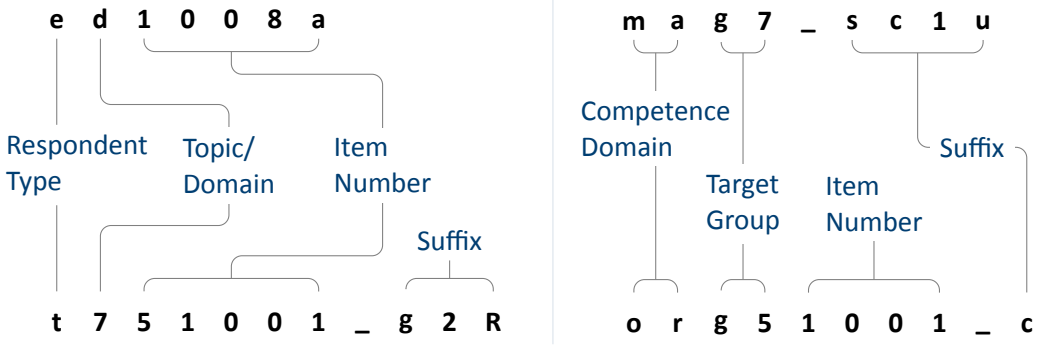

Figure 5: General variable naming (left) and competence variable naming (right)

3.2.1 Conventions for general variable naming

A general NEPS variable name consists of up to four elements – the respondent type, the domain of information, an item number, and one or more optional suffixes.

Table 4: Conventions for variable names

Digit	Description
1	**Respondent type** Indicator to which group of respondents the variable refers; please note that variables related to the target person start with t even if the target person was not the actual informant (e. g., list data from schools/kindergartens) t = Target person p = Parent/legal guardian of target person c = Cohabiting partner of the target child's parent e = Educator/childminder/teacher h = Head/administration of school or 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. Since scale indices are generated by a set of other variables, they are also identified by a `_g#` suffix. It is important to know 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 (e. g., `_g1`). However, there are three types of generated variables that assign specific meanings to digits, namely regional classifications, occupational variables, and education variables.

The **regional classifications** are based on the Nomenclature of Territorial Units for Statistics (NUTS) and refer to units within Germany:

- 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 **occupation and prestige indices** are (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)

Generated variables for the **classification of highest educational qualification** are:

- g1: ISCED-97 (Int. Standard Classification of Education 1997, *not in Starting Cohort 8*)
- g2: CASMIN (Comparative Analysis of Social Mobility in Industrial Nations)
- g3: Years of Education (derived from CASMIN)
- g4: ISCED-2011 (International Standard Classification of Education 2011)

- *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 is 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 typically 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. **Please be aware that this suffix does not necessarily imply a wave logic.** The presence of a set of variables with _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, and _O suffixes indicate 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 (_R) or in combination with other suffixes (_g3R).

3.2.2 Conventions for competence variable naming

The naming of variables from competence measurements and direct measures follows a different logic. In contrast to other data files, the competence datasets 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.[7]

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

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 from IQB-BT
md	Declarative metacognition
mp	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 from IQB-BT
re	Reading competence
ri	Rimes: Phonological awareness
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

(...)

Table 5: (continued)

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

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

Part III: Item number (3-4 chars)

For most competence domains, these item numbers only indicate 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)
_sc8	Test stop
_sc9	Basal/Ceiling set (for vocabulary), number of practice items (for digit span)
_p	Maximum value for an item (only in xDirectMeasures of Starting Cohort 1)
_b	Minimum value for an item (only in xDirectMeasures of Starting Cohort 1)
_m	Mean value for an item (only in xDirectMeasures of Starting Cohort 1)
_s	Sum value for an item (only in xDirectMeasures of Starting Cohort 1)
_n	Number value for an item (only in xDirectMeasures of 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).

The additional suffixes inform about the mode of test execution if more than one survey modus has been applied for a measurement and about the sort of item score and overall competence score. There is a distinction between scored items named [varname]_c and scored partial-credit items named [varname]s_c. The latter is relevant if more than one correct solution is possible (e. g., value 0 = “0 out of two points”, value 1 = “1 out of two points”, value 2 = “2 out of

two points"), whereas the former is applied for dichotomous solutions (value 0 = "not solved", value 1 = "solved"). In addition to the 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 lists all possible suffixes in competence variable names and their meanings.

Identification of repeated test items

For some competence measurements, there were identical items implemented in different panel waves (e. g., mathematics). Identifying repeatedly measured test items in the NEPS data can easily be done by looking for competence variables with an identical word stem. If the same test item has been 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 a repeated use consists of two parts: The first element indicates the starting cohort of current item administration and the second element indicates the time of current item administration.

The following example illustrates this logic: The competence variable `vok10067_sc2g1_c` is a vocabulary item (`vo`), initially measured during the first kindergarten survey wave of Starting Cohort 2 (`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

The seven-digit variable names are not sufficient to completely identify the respective contents of the variables and to differentiate between certain items. Therefore, all variables have *variable labels* for a more detailed description of what they entail. Most variables also 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).

In addition to the variable and value labels, the datasets also feature so-called "extended characteristics". They include the question text of an item from the survey instrument, associated

filter conditions, as well as other meta information. All extended characteristics can be accessed directly within the datasets. 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.[8]

As explained in more detail in Section 4, the data from different waves is integrated as much as possible. For panel data, this primarily means that variables contain information from multiple waves. In most cases of such an integration, the meta information between the waves does not change. However, if there are changes in the meta information of a repeatedly measured item, and if these changes are not significant enough to store the data in separate variables, the assignment 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 change 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 are integrated into one variable. If the meta information of such a variable in the dataset is requested, the wording of the latest item formulation will be displayed (in the given example the formal salutation "Sie"). In case of uncertainties with regard to the continuity of meta information of a certain variable across waves, it is recommended to consult the original survey instruments for the respective survey waves.

3.3 Missing values

There are several missing codes in the NEPS data to differentiate between types of missing values. All missing codes have negative values or are defined as "system missing". Depending on the statistics program used for analyzing the data, one has to ensure that these codes are processed correctly. In the SPSS datasets, the missing codes are already defined as missing values. When using Stata, the missing values must first be defined or excluded from the analysis. 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 variables before running an analysis. The three main types of missing codes are summarized in Table 6 and described below.

Due to the survey software used in Starting Cohort 8, nearly all system missing values (.) were recoded retrospectively into three meaningful categories of missings as part of the data editing process for the Scientific Use File (see Section 5.7 for details).

8 Some variables are based on different versions of a question formulation, depending on previous filter settings (e.g., stay abroad of at least one month for training/ for study/ for doctorate etc. [ts15223 in SC3, SC4, SC6] according to the current type of training of the respondent [ts15223]). Only the first version is provided in the extended characteristics, but an additional note is displayed: "ATTENTION: There are several question formulations for this variable – see the Variable Search or the Survey Instruments for detailed information!"

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 the 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)
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) or missing information despite soft response constraint (CAWI/CASI)
–91	survey aborted	question not reached due to early termination of the instrument (CAWI/CASI)
–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)
–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.

- Typical cases of item nonresponse are "refused" (–97) answers and "don't know" (–98) answers to questions.
- 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 when 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 either the respondent 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 (see Section 5.7 for additional information); in all other modes the respective code is "filtered" (–99).
- If a respondent has terminated the survey in an online mode (CAWI/CASI) before reaching the end of the instrument or if the survey session has been aborted by timeout, the missing code "survey aborted" (–91) is assigned to all questions that were not answered at the end of the questionnaire.
- 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, or in CAWI/CASI interviews when a respondent has refused to answer a question despite soft response constraint.

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

- If during the data edition 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. In general, however, all values from the source data 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 code is inserted.
- 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).
- If information from the source data is not sufficient to derive a variable (e. g., occupational categories; see Section 3.4), the missing code "not determinable" (–55) is used.
- 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. It 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 freely written text. A typical example is the job title.

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 input 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 procedure applies to semi-open question formats where respondents enter a text under the category "other", 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 FDZ-LIfBi itself. Other coding tasks are distributed among the responsible working units at the LIfBi in Bamberg and the responsible 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 in a user-friendly and analyzable way within the NEPS Scientific Use File. 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 variable 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 6 shows the derivation paths for several **occupational scales and schemes** provided in the NEPS.

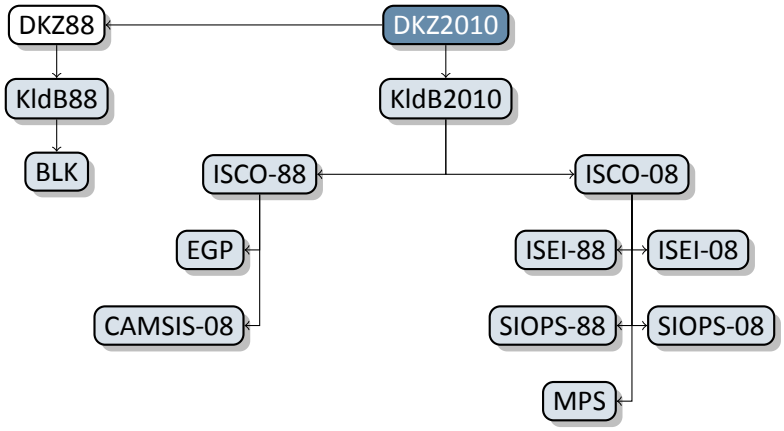

Figure 6: Derivation paths for several occupational scales and schemes provided in the NEPS

A description of the standard derivations for **educational attainment** (ISCED, CASMIN and Years of Education) can be found in the corresponding documentation report by Pelz, 2025.

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.

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. 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).[9]

This convention does not apply to all longitudinal information in the Scientific Use File. 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. For the information collected prospectively and retrospectively by using iterative question sets, the Scientific Use File provides life area-specific spell datasets. They are marked by a preceding ***sp***. An example is `spEmp`, informing about current and former episodes of employment (see Section 4.4).[10]

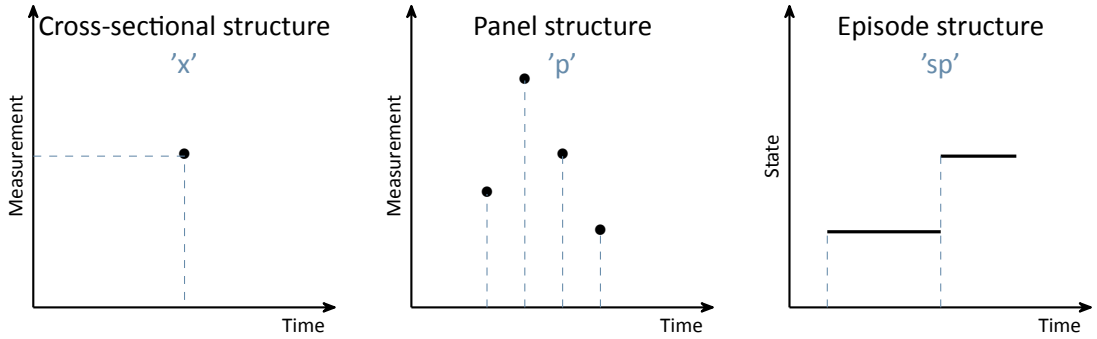

Figure 7: Different types of data structures

In addition to the interview and competence and episode data surveyed from the respondents, there are so-called paradata and derived information available. The respective data files can be

9 Even though the Scientific Use File for Starting Cohort 8 currently only contains information from the first survey wave, the prospective panel datasets are already marked with the ***p*** in the file name.
10 In the first two survey waves of Starting Cohort 8, no episode data were collected from the respondents.

identified by the leading capital letter in the name (e. g., `CohortProfile`, `TargetMethods` or `Weights`; see Figure 8).

4.2 Identifiers

The multi-level and multi-informant design of the NEPS – together with the provision of information in several data files – requires the use of multiple identifiers, especially for merging information from different datasets. The syntax examples in Section 4.5 demonstrate typical applications for linking information from the respective file to information from other files. The following identifier variables are particularly relevant in Starting Cohort 8:

ID_t identifies a target person persistently. The variable `ID_t` is unique across waves and samples; it is also used uniquely in each of the NEPS starting cohorts.

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

ID_p identifies the surveyed parent of a target person persistently. If the interviewed parent changes during the course of the panel survey, the identifier `ID_p` changes accordingly.[11]

ID_i identifies the educational institutions attended by the target person (e. g., kindergartesn or day care centers, schools, universities). The variable `ID_i` is unique across waves and starting cohorts.

ID_e identifies surveyed educators or teachers persistently. This identifier variable can be used to merge data from such context persons with observations from the children. **However, it is not possible to merge the data directly with `ID_t`.** The linking with data of the target persons requires the "path" via the dataset `LinkTargetTeacher` that is offered specifically for this purpose. An example is given in the respective dataset description (see Section 4.5.3).

ID_cc identifies the class that the target person attends – within a wave. This identifier variable is **not** unique across waves.

There are further identifier variables available, for example 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.

11 The persistence of the parent ID across the survey waves is a major innovation of Starting Cohort 8 compared to the previous NEPS starting cohorts.

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 NEPS 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 rather 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).[12]

In episode data, there is one row for each episode that was captured during an 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

12 This data type does not yet play a role in the current Scientific Use File of Starting Cohort 8.

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.

4.5 Data files

In the following section, every data file of Starting Cohort 8 is described in a subsection, including a data snapshot and a syntax example that often deals with the challenge of merging information from another file (in Stata). The syntax examples are written in an easily comprehensible way. There is no need to additional install any “ado files”, although it is highly advised to use the `NEPStools` (see Section 1.6).

To facilitate the understanding of the relationships between the data files in the Scientific Use File, an overview of all datasets is provided in Figure 8. The lines in this figure symbolize how a data file may be linked to other files. This is not meant to document every possible data link, but rather tries to give an idea on which data files relate most. By clicking on a box, one gets directed to the short description of this dataset.

For the Stata syntax examples in the subsequent dataset short portraits to work, the following globals must first be set. Just adapt and copy the lines below to the top of the syntax files or execute them in the Stata command line before running the syntax.

```
** Starting Cohort
global cohort SC8
** version of this Scientific Use File
global version 2-0-0
** path where the data can be found on your local computer
global datapath Z:/Data/${cohort}/${version}
```

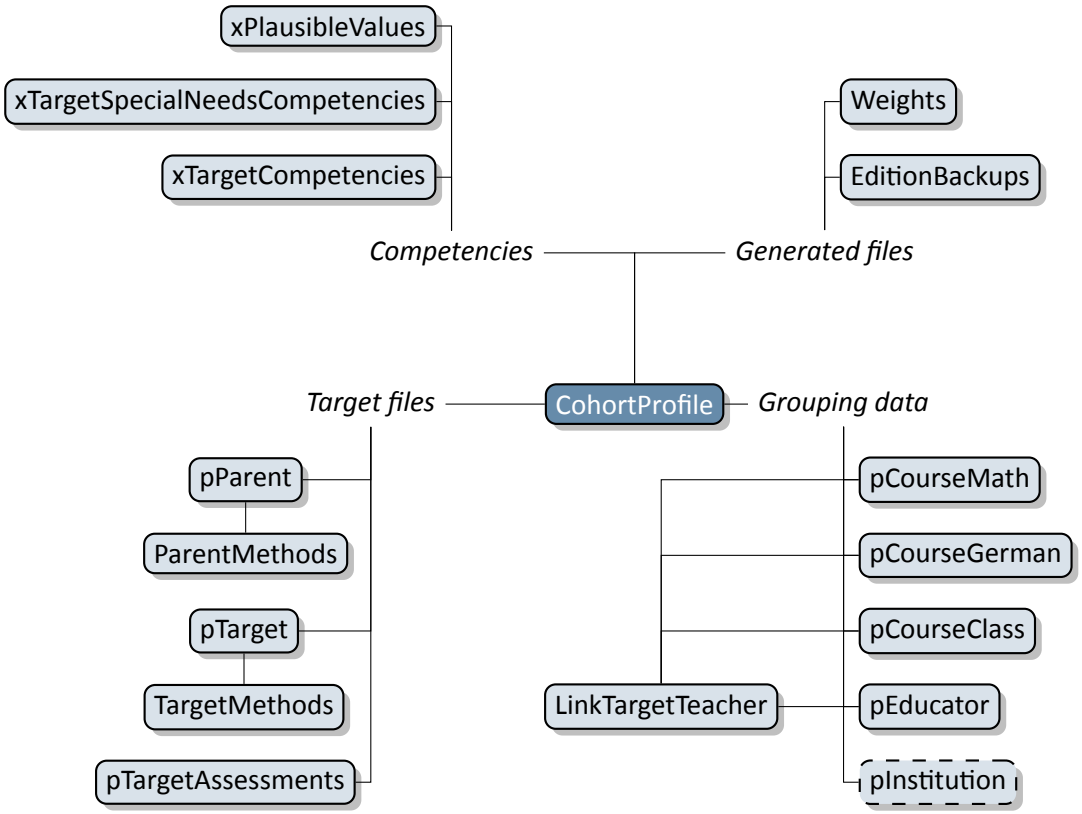

Figure 8: Graphical overview of all data files. Each box represents one data file. Relations are indicated by connecting lines. Files with a dashed border are not available in the Download version of the Scientific Use File. Click on a file to get more information.

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_i ID_cc ID_tg

Number of variables / number of rows in file

64 / 12,282

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
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
tx8600y	Survey Target person: survey year
tx80524	Data available: institution

Exemplary data snapshot

ID_t	wave	tx80220	tx80521	tx80522	tx8610y	tx8600y	tx80524
4087782	1	Participation	yes	yes	2022	2022	yes
4088508	1	Participation	yes	yes	2023	2023	yes
4089262	1	Participation	yes	yes	2023	2023	yes
4090042	1	Participation	yes	yes	2022	2022	yes
4090917	1	Participation	yes	yes	2022	2022	yes

The dataset `CohortProfile` includes all target children of the panel sample. These are all students with an initial agreement to participate. This dataset is particularly suitable for selecting cases for the analysis and for merging information from different datasets.[13]

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

For each student and each wave, the `CohortProfile` contains relevant ID variables (person: `ID_t`, class: `ID_cc`, school: `ID_i`), but also meta information such as the type of school according to the sampling frame (`tx80106`), the participation status (`tx80220`), the availability of specific data (e. g., competence data `tx80522`), the school grade (`tx80234`), or the gender and birth date of the target children according to the list of students (`tx80501/tx8050*`). In addition, there are variables of the dates when the competence tests and interviews took place (`tx8610*/tx8600*/tx8620*`). Please note, that all dates are provided in months and years only.

13 Another dataset made available for merging purposes is the `LinkTargetTeacher` file.

Stata 1: Working with CohortProfile

```
** open the data file
use "${datapath}/SC8_CohortProfile_D_${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_cc wave

Other ID variables useful for linkage

mergevars

Number of variables / number of rows in file

11 / 264

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
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
4089843	1	pTarget	t400030	ID_t wave	..00	5.00
.	1	pEducator	e76212m	ID_e wave	10.20	1.00
.	1	pEducator	e76212m	ID_e wave	1.20	1.00
.	1	pEducator	e76212m	ID_e wave	20.10	1.00
.	1	pInstitution	h227000	ID_e wave	..00	100.00

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
** t741002 (Household size) of datafile pTarget

** open the datafile
use "${datapath}/SC8_EditionBackups_D_${version}.dta", clear

** only keep rows containing data of the aforesaid variable
keep if dataset=="pTarget" & varname=="t741002"

** check which variables we need for merging
tab mergevars

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

** rename the variables to emphasize affiliation
rename sourcevalue_num t741002_source
rename editvalue_num t741002_edit

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

** open pTarget
use "${datapath}/SC8_pTarget_D_${version}.dta", clear

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

** check all edition made
list ID_t wave t741002* if _merge==3

** replace the variable in the datafile with its original value
replace t741002=t741002_source if _merge==3
```

4.5.3 LinkTargetTeacher

« go back to overview

Description

Combination of ID variables on schools, classes, teachers, students

File structure

M:N format: 1 row = 1 combination of student ID and teacher ID per wave

ID variables needed to identify a single row

ID_t ID_e wave

Other ID variables useful for linkage

ID_cc ID_i tx20103

Number of variables / number of rows in file

6 / 12,591

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave	Wave
tx20103	Teacher specifications per student
ID_e	ID teacher/educator
ID_cc	Course-ID: Grade
ID_i	Institution ID

Exemplary data snapshot

ID_t	wave	tx20103	ID_e	ID_cc	ID_i
4086078	1	5	1020743	10051581002	1005158
4086078	1	5	1021754	10051581002	1005158
4086078	1	5	1022686	10051581002	1005158
4086078	1	5	1023356	10051581002	1005158
4086078	1	5	1023454	10051581002	1005158
4086078	2	1	1022686	10051582002	1005158
4086183	1	3	1022661	10050461001	1005046
4086183	1	3	1023030	10050461001	1005046
4086183	1	3	1023222	10050461001	1005046
4086183	2	2	1023030	10050462001	1005046
4086183	2	2	1023222	10050462001	1005046

The `LinkTargetTeacher` dataset was generated to represent all possible ID combinations of students, teachers, classes and schools. Due to the survey design of Starting Cohort 8, a simple linkage is generally not possible due to the m:n data structure – there are usually several classes per school included in the survey; each class usually has several students in the sample; and there are often responses from several teachers for one single class.

It is therefore important to first think about the type of analysis and the data structure required for this before merging information from different datasets.

One can either exclude or aggregate certain information (e. g. by selecting only one teacher statement for a certain class or by averaging a certain characteristic from the various statements of several teachers per class) and then use these reduced data rows as the basis for adding the desired information from other datasets. Or, one can use the various ID variables in their multiple structure in this dataset to first retrieve information from other datasets (e. g. 1:m merge via `ID_t` with `pTarget` or via `ID_e` with `pEducator`) and then make a selection of the relevant cases or generate the data structure required for the analysis via aggregations etc.

Stata 3: Working with LinkTargetTeacher

```
** this example illustrates how you can add specific information from various other
** sources to LinkTargetTeacher

** open the LinkTargetTeacher file
use "${datapath}/SC8_LinkTargetTeacher_D_${version}.dta", clear

** append respondent information to this file (e.g., gender from CohortProfile)
merge m:1 ID_t wave using "${datapath}/SC8_CohortProfile_D_${version}.dta", nogen
  keep(master match) keepusing(tx80501)

** append teacher information (e.g., gender from pEducator)
merge m:1 ID_e wave using "${datapath}/SC8_pEducator_D_${version}.dta", nogen keep(
  master match) keepusing(e762111)

** append class information (e.g., team teaching from pCourseClass)
merge m:1 ID_cc ID_e wave using "${datapath}/SC8_pCourseClass_D_${version}.dta",
  nogen keep(master match) keepusing(ed0604f)
```

Stata 4: Additional example with LinkTargetTeacher

```
** in this example, we add the teacher ID (ID_e) to the pTarget datafile,
** so additional information from pEducator can be merged

use "${datapath}/SC8_LinkTargetTeacher_D_$version.dta", clear

** reduce to one teacher information per class - take first teacher
bysort ID_t wave ID_cc (ID_e): gen teachernumber = _n
keep if teachernumber == 1
drop teachernumber

isid ID_t
** --> dataset now is unique with respect to ID_t and can be merged to pTarget

** save a temporary file
tempfile link
save `link'

use "${datapath}/SC8_pTarget_D_$version.dta", clear

** merge ID_e from linkfile
merge 1:1 ID_t wave using `link', keep(master match) nogen keepusing(ID_e)
```

```
** add teacher information from pEducator
merge m:1 ID_e wave using "${datapath}/SC8_pEducator_D_$version.dta", keepusing(
  e76212y_R) nogen keep(master match)
```

Stata 5: Additional example with LinkTargetTeacher

```
** this examples shows how you could compare the age of teachers
** (note that this needs at least Remote version of SUF)

** open LinkTargetTeacher ...
use "${datapath}/SC8_LinkTargetTeacher_R_$version.dta", clear

** ... and merge survey year and date of birth
merge m:1 ID_e wave using "${datapath}/SC8_pEducator_R_$version.dta", nogen keep(
  master match) keepusing(e76212y_R ex8601y)

** generate approximate teacher age (additionally using months would be more precise)
gen teacherage = ex8601y - e76212y_R if e76212y_R >0 & ex8601y >0

** generate mean age of all teachers per student per wave
bysort ID_t wave (ID_e): egen teacherage_mean = mean(teacherage)

** recode into 10 year classes
recode teacherage_mean (20/30 =1 "21-30") (30/40 = 2 "31-40") (40/50 = 3 "41-50")
  (50/60 = 4 "51-60") (60/70 = 5 "61-70") (70/max = 6 "older than 70"), gen(tagemean)
label variable tagemean "mean age of all teachers"

** keep only variables of interest
keep ID_t wave tagemean

** drop unneccessary duplicates
duplicates drop

** save a temporary file ...
tempfile link
save `link'

** .. and merge this to pTarget
use "${datapath}/SC8_pTarget_R_$version.dta", clear
merge 1:1 ID_t wave using `link', keep(master match) nogen

** show distribution of teacher age mean
fre tagemean
```

4.5.4 ParentMethods

« go back to overview

Description

Paradata from the parents CATI interview

File structure

Long format: 1 row = 1 parent in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

ID_p ID_int

Number of variables / number of rows in file

36 / 6,978

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave	Wave
px80200	Interview: number of all contact attempts
px80209	Interview: length of interview (minutes)
px80212	Interview: change of contact person to previous wave
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
4086477	1	53.36666	3583	male	30-49 years
4087080	1	55.99166	3583	male	30-49 years
4089380	1	47.63334	3162	female	50-65 years
4089574	1	58.87778	3589	male	older than 65 years
4090752	1	46.86111	1039	female	50-65 years

This dataset offers information on the data collection during the interview with the surveyed parents. These include characteristics of the interviewer such as gender (px80301) and age (px80302), the survey mode (px80202), the interview length (px80209), the differentiated response code (px80207), assessments of the interviewer on the interview (e. g., reliability px80323) etc.

Please note that this file contains all contacted parents, whether an interview was realized or not. Thus, ParentMethods includes more cases than the data file pParent. **It is also important to know that the parent ID (ID_p) is persistent across waves.**

Stata 6: Working with ParentMethods

```
** open the data file
use "${datapath}/SC8_ParentMethods_D_${version}.dta", clear

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

** check out response code by wave
tab px80207 wave

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

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

4.5.5 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

Other ID variables useful for linkage

ID_i ex20102

Number of variables / number of rows in file

80 / 426

Contains data from waves

1 2

Exemplary variables

ID_cc	Course-ID: Grade
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
ex20102	Teacher specifications per class
e410011_g1R	First language/mother tongue Educator (ISO 639.2)
e79204a_R	Social composition of class: education
ed11500	Experience inclusion - in general

Exemplary data snapshot

ID_cc	wave	ID_e	ID_i
10052091002	1	1020840	1005209
10050961002	1	1020865	1005096
10051321001	1	1021971	1005132
10050381003	1	1021816	1005038
10050301002	1	1023346	1005030

This data file contains the information collected from the class teachers about the respective classes. This concerns information on the social composition of the class (`e79204*_R`), on teacher stereotypes (`e31605*/e31606*`), on co-teaching (`ed0604*`), on teaching and class climate (`ed1104*`), and on special needs issues (`ed1152*_R`). The educators reporting these information can be identified via `ID_e`, the respective school via `ID_i`.

In several cases, more than one educator reported information about a single class. The variable `ex20102` indicates the number of teachers that answered questions on the same class.

To combine information from this data file with information from other data files, it is important to first consider the necessary data structure for the intended analysis. In many scenarios it makes sense to remove or aggregate information from teachers on the same class before merging the data (see also the description of `LinkTargetTeacher`).

Stata 7: Working with pCourseClass

```
** open the data file pCourseClass
use "${datapath}/SC8_pCourseClass_D_${version}.dta", clear

** be aware that for each class, multiple rows are existent.
** this is because the instrument has beeen reported by multiple educators
duplicates report ID_cc wave

** to further work with this data, we reduce it to one single row
** for each class. We do this by just keeping the first observation for a class.
** note that in a real situation, you most likely have to go into more detail here,
** and evaluate what to keep and how to do this
bysort ID_cc wave: keep if _n==1

** save this temporarily ...
tempfile pCourseClass
save `pCourseClass'

** ... and merge it to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge m:1 ID_cc wave using `pCourseClass', nogen keep(master match)
label language en

** now view how many respondents experience team teaching
** (ed0604f: Co-teaching form - team teaching)
tab tx80521 ed0604f
```

4.5.6 pCourseGerman

« go back to overview

Description

Data about teaching in German classes

File structure

Long format: 1 row = 1 German class in 1 wave

ID variables needed to identify a single row

ID_cc ID_e wave

Other ID variables useful for linkage

ID_i ex20102

Number of variables / number of rows in file

65 / 257

Contains data from waves

1 2

Exemplary variables

ID_cc	Course-ID: Grade
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
ex20102	Teacher specifications per class
e22345a	Teaching quality - class management: class rules
e22686c	Self-efficacy - teaching: good questions
ed0750c	Forms of learning/organization: class teaching
e22141a	Teaching quality - frequency digital media German: research

Exemplary data snapshot

ID_cc	wave	ID_e	e22345a
10051161001	1	1021871	often
10052281002	1	1021828	very often
10052251005	1	1022035	often
10051401001	1	1021126	often
10051841003	1	1022518	very often

This data file contains the information collected from the German teachers about the respective courses. This concerns information on teaching quality (`e22141*-545*`), on self-efficacy (`e22686*`), on teacher cooperation (`e23207*`), on praise strategies (`ed0700*`), and on forms of learning/organization (`ed0750*`). The educators reporting these information can be identified via `ID_e`, the respective school via `ID_i`.

In several cases, more than one educator reported information about a single class. The variable `ex20102` indicates the number of teachers that answered questions on the same class.

To combine information from this data file with information from other data files, it is important to first consider the necessary data structure for the intended analysis. In many scenarios it makes sense to remove or aggregate information from teachers on the same class before merging the data (see also the description of `LinkTargetTeacher`).

Stata 8: Working with pCourseGerman

```
** open the data file pCourseGerman
use "${datapath}/SC8_pCourseGerman_D_${version}.dta", clear

** be aware that for each class, multiple rows are existent.
** this is because the instrument has beeen reported by multiple educators
duplicates report ID_cc wave

** to further work with this data, we reduce it to one single row
** for each class. We do this by just keeping the first observation for a class.
** note that in a real situation, you most likely have to go into more detail here,
** and evaluate what to keep and how to do this
bysort ID_cc wave: keep if _n==1

** save this temporarily ...
tempfile pCourseGerman
save `pCourseGerman'

** ... and merge it to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge m:1 ID_cc wave using `pCourseGerman', nogen keep(master match)
label language en

** now view how many respondents experience small group work
** (ed0750a Forms of learning/organization: small group work)
tab ed0750a tx80521
```

4.5.7 pCourseMath

« go back to overview

Description

Data about teaching in math classes

File structure

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

ID variables needed to identify a single row

ID_cc ID_e wave

Other ID variables useful for linkage

ID_i ex20102

Number of variables / number of rows in file

64 / 230

Contains data from waves

1 2

Exemplary variables

ID_cc	Course-ID: Grade
wave	Wave
ID_e	ID teacher/educator
ID_i	Institution ID
ex20102	Teacher specifications per class
e22346a	Teaching quality - class management: class rules
ed0850h	Forms of learning/organization: project learning
e22142a	Teaching quality - frequency digital media math: research
e22247e	Teaching quality - support digital media math: work on tasks

Exemplary data snapshot

ID_cc	wave	ID_e	ID_i	ed0850h
10051741001	1	1023037	1005174	a few times a year
10051931002	1	1021249	1005193	never
10051801004	1	1022858	1005180	a few times a month
10050601001	1	1021172	1005060	never
10050241002	1	1021874	1005024	a few times a month

This data file contains the information collected from the math teachers about the respective courses. This concerns information on teaching quality (`e22142*-546*`), on self-efficacy (`e22688*`), on teacher cooperation (`e23208*`), on praise strategies (`ed0800*`), and on forms of learning/organization (`ed0850*`). The educators reporting these information can be identified via `ID_e`, the respective school via `ID_i`.

In several cases, more than one educator reported information about a single class. The variable `ex20102` indicates the number of teachers that answered questions on the same class.

To combine information from this data file with information from other data files, it is important to first consider the necessary data structure for the intended analysis. In many scenarios it makes sense to remove or aggregate information from teachers on the same class before merging the data (see also the description of `LinkTargetTeacher`).

Stata 9: Working with pCourseMath

```
** open the data file pCourseMath
use "${datapath}/SC8_pCourseMath_D_${version}.dta", clear

** be aware that for each class, multiple rows are existent.
** this is because the instrument has beeen reported by multiple educators
duplicates report ID_cc wave

** to further work with this data, we reduce it to one single row
** for each class. We do this by just keeping the first observation for a class.
** note that in a real situation, you most likely have to go into more detail here,
** and evaluate what to keep and how to do this
bysort ID_cc wave: keep if _n==1

** save this temporarily ...
tempfile pCourseMath
save `pCourseMath'

** ... and merge it to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge m:1 ID_cc wave using `pCourseMath', nogen keep(master match)
label language en

** now view how many respondents experience small group work
** (ed0850a Forms of learning/organization: small group work)
tab ed0850a tx80521
```

4.5.8 pEducator

« go back to overview

Description

General information from the teachers

File structure

Long format: 1 row = 1 educator in 1 wave

ID variables needed to identify a single row

ID_e wave

Other ID variables useful for linkage

ID_i

Number of variables / number of rows in file

211 / 2,814

Contains data from waves

1 2

Exemplary variables

ID_e	ID teacher/educator
wave	Wave
e400000	Migrant background
e514001	General satisfaction: life
ed1101a	Job satisfaction: work meaningful and important
ed1102c	Occupational stress: satisfied
e53716a	Type teaching degree course: elementary school
e537167_g1	Education: type of qualification (categorized)
ed0605a	Attitudes: improve education
e222000	Media-didactic concept: existence concept

Exemplary data snapshot

ID_e	wave	e400000	e537167_g1
1023205	1	3	Diploma, Master (M.A., M.Sc.)
1023006	1	3	other vocational qualification
1023323	1	3	Diploma, Master (M.A., M.Sc.)
1023535	1	3	other vocational qualification
1021465	1	3	Magister, state examination

In addition to the class or course-specific modules, all surveyed teachers answered a general part of the questionnaire. The information from this module is stored in `pEducator`. The contents relate, for example, to ratings on school culture (`e22208*`), on dealing with diversity in school life (`e22222*`), on teaching evaluation (`e2228*`), on aspects of leadership (`e22237*`), on the use of digital media (`e22244*`) and the promotion of IT skills (`e22245*`). Other content refers to the person of the teacher such as gender (`e762111`) and birth date (`e76212*`), immigrant background (`e400000`), teaching qualifications (`e53718*`), working time (`e229822/e229823`), and several attitudes (e. g., job satisfaction `ed1101*`).

This file contains all educators from the sample, no matter if they were class teachers, German teachers or math teachers. For the first time in NEPS, the study design included an invitation to all teachers at the participating schools, not just addressing those teaching the selected NEPS classes.

Please note that teacher and student information can be merged, but there is no direct link between teacher IDs and student IDs via `CohortProfile`. This is due the following reasons:

- By study design, one student could have several teachers (class, German, math).
- In some cases, more than one class/German/math teacher answered the respective questions on a single class or course (see variable `ex20102`) in the respective datasets.
- There is no 1:1 relationship in the design between teachers and students. Teachers have only been interviewed about themselves, as well as about the class. There were no questions asked to the teachers about individual students in wave 1.

To combine information from this data file with information from other data files, it is important to first consider the necessary data structure for the intended analysis. In many scenarios it makes sense to remove or aggregate information from teachers on the same class before merging the data (see also the description of `LinkTargetTeacher`).

Stata 10: Working with pEducator

```
** open the LinkTargetTeacher file
use "${datapath}/SC8_LinkTargetTeacher_D_${version}.dta", clear

** and merge pEducator
merge m:1 ID_e wave using "${datapath}/SC8_pEducator_D_${version}.dta", nogen keep(
  match)

** e76212y_D : Date of birth: year (categorized)
keep ID_t wave ID_e ID_cc e76212y_D


** ... now proceed as in the example of LinkTargetTeacher
bysort ID_t wave ID_cc (ID_e): gen teachernumber=_n
drop ID_e ID_cc
reshape wide e76212y_D, i(ID_t wave) j(teachernumber)

** save this temporarily ...
tempfile pEducator
save `pEducator'

** ... and merge it to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge 1:1 ID_t wave using `pEducator', nogen keep(master match)
label language en

** now have a look at the age distribution of teachers
sort e76212y_D5
list ID_t e76212y_D1 e76212y_D2 e76212y_D3 e76212y_D4 e76212y_D5 in 1/10
```

4.5.9 pInstitution

« go back to overview

Description

Context data collected from school management (principals)

File structure

Long format: 1 row = 1 principal or school management person information in 1 wave

ID variables needed to identify a single row

ID_e wave

Other ID variables useful for linkage

ID_i

Number of variables / number of rows in file

186 / 633

Contains data from waves

1 2

Exemplary variables

ID_i	Institution ID
wave	Wave
hd0060a	Job satisfaction: work meaningful and important
hd00580	School profile: available
hd0101a	Type of special needs: learning
hd00580	School profile: available
hd0059c	School profile: new languages
hd0059d	School profile: humanistic ancient languages
hd01040	Full-day program
hd0101a	Type of special needs: learning

Exemplary data snapshot

ID_i	wave	hd0060a	hd00580	hd0059c
1005196	1	very often	1	specified
1004977	1	very often	1	specified
1005018	1	very often	1	specified
1005167	1	very often	1	specified
1005167	1	very often	1	specified

This data file contains information collected from persons of the school management about the school. This concerns, for example, issues of school culture (h22208*/09*), the curriculum (h22224*), the pedagogical concept (h229030), extracurricular activities (h34021*-031*), school challenges (hd0053*) and the school profile (hd0059*), but also the composition of teaching staff and student body with regard to immigrant background (h451080/h451020). Other content refers to personal characteristics such as gender (h766111) and birth date (h76612*) as well as attitudes such as job satisfaction (hd0060*), occupational stress (hd0061*), and the importance of performance (hd0064*).

This dataset is not available in the Download version of the Scientific Use File.

In several cases, more than one person reported information about a single institution/school. The variable ID_e differentiates between the various informants within one school (ID_i).

To combine information from this data file with information from other data files, it is important to first consider the necessary data structure for the intended analysis. In many scenarios it makes sense to remove or aggregate information from informants on the same school before merging the data.

Stata 11: Working with pInstitution

```
** open the pInstitution file
** note that we operate on remote (_R)
use "${datapath}/SC8_pInstitution_R_${version}.dta", clear

** notice that multiple rows exist per school
** (the same questionnaire has been filled out by multiple educators/"headmasters")
duplicates report ID_i wave

** we reduce this to one row for each school.
** in this example, we only keep the data from the
** longest serving educator (h229823_R: Time working in profession: school)
bysort ID_i wave (h229823_R): keep if _n==_N

** save this temporarily ...
tempfile pInstitution
save `pInstitution'

** ... and merge it to CohortProfile
use "${datapath}/SC8_CohortProfile_R_${version}.dta", clear
merge m:1 ID_i wave using `pInstitution', nogen keep(master match)
label language en

** You now have school-dependent information for each respondent, e.g.
** check if for students in schools with the profile
** hd0059i: "School profile: theater/performing artist"
** parents have been surveyed
tab hd0059i tx80523
```

4.5.10 pParent

« go back to overview

Description

Data surveyed from parents/legal guardians

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_p

Number of variables / number of rows in file

509 / 5,648

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave	Wave
p731905	Professional position respondent
p731955	Professional position partner
p731701	Relationship respondent to target child
p741001	Household size
p400500_g1	Generation status
p73170y	Year of birth respondent
p731835	Course of study abroad respondent
p731912	Management position respondent

Exemplary data snapshot

ID_t	wave	p731905	p731955	p731701	p741001	p400500_g1
4086074	1	3	2	biological mother	5	10
4086451	1	2	2	biological mother	4	2
4088463	1	2	2	biological mother	4	10
4088824	1	7	5	biological father	5	10
4090051	1	2	2	biological mother	4	2

This data file contains the information collected from the parents or legal guardians of the target children. The topics surveyed cover personal characteristics of the responding person and the partner such as gender (p731707/p731122), year of birth (p73170y/p73175y), country of birth (p400000/p403000), marital status (p731110), educational qualification (p7318*), occupation (p7319*), but also health-related issues (e. g., smoking p5250*), political engagement (e. g., interest p516105) and cicic participation (e. g., memberships p51760*). Other topics are related to the household level such as size (p741001), monthly income (p510005), material deprivation (p74310*), and interaction language (p41203b). However, the majority of information relates to the target child: from birth (e. g., size and weight p52900*) and early childhood (e. g., breastfeeding p52620*) to current issues (e. g., self-rated health p521000) and to education-specific topics such as language learning (e. g., age at start of learning German p41002*), special educational needs (e. g., diagnosis pd01025), recommendation after elementary school (pd01016), parent-teacher cooperation (p221874) and various aspirations (e. g., idealistic highest school-leavaing qualification p31035a).

Stata 12: Working with pParent

```
** open the CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear

** merge occupation of parents (both respondent and partner) from pParent
merge 1:1 ID_t wave using "${datapath}/SC8_pParent_D_${version}.dta", ///
  keepusing(p731905 p731955) nogen assert(master match)

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

** recode missings
nepsmiss p731905 p731955

** 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 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

none

Number of variables / number of rows in file

497 / 10,663

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave	Wave
t700032	Sex at birth
t70004y	Date of birth: year
t400000_g1R	Country of birth
t741002	Household size
t34007a	Cultural participation: high culture
t214211	Use of digital media at school: German
t22344a	Teaching quality - discipline: listening to teacher
t435000	Religion & religiousness: religiousness
t34005a	Number books

Exemplary data snapshot

ID_t	wave	t70004y	t741002	t34007a	t214211
4088072	1	2012	4	2 to 3 times	in some lessons
4089373	1	2012	4	once	in some lessons
4091127	1	2011	4	2 to 3 times	in some lessons
4092083	1	2012	7	more than 5 times	in some lessons
4096544	2	2011	3	never	in some lessons

This data file contains the information collected from the target children on the basis of questionnaires. The topics surveyed cover personal characteristics such as gender (`t700033`), date of birth (`t70004*`), country of birth (`t400000*`), mother tongue (`t414000*`), sense of belonging (`t4280*`) and extracurricular activities (`t26162*`), but also household-related issues such as composition (`t74305*`). However, the majority of information relates to education-specific topics such as self-concept (`t6600*`), idealistic and realistic aspirations (e. g., highest school-leavaing qualification `t31035a/f`), tutoring (`t26111*`), use of digital media at school (`t2142*`), teaching quality (`t22344*`), motivation (`t66409*`) and parental involvement (`t28165*-269*`).

Stata 13: Working with pTarget

```
** open the CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear


** merge country of birth and generation status from pTarget
merge 1:1 ID_t wave using "${datapath}/SC8_pTarget_D_${version}.dta", ///
  keepusing(t400500_g1 t400000_g1D) nogen

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

** recode missings
nepsmiss t400500_g1 t400000_g1D

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

4.5.12 pTargetAssessments

« go back to overview

Description

Data from the class teacher's individual assessment of the target children

File structure

Long format: 1 row = 1 assessment by 1 class teacher of 1 target child in 1 wave

ID variables needed to identify a single row

ID_t ID_e wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

21 / 2,126

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
ID_e	ID teacher/educator
wave	Wave
ed1105a	Subject mathematics
ed1105b	Subject German
e66013a	Assessment: diligence
e66013b	Assessment: social behavior
e66013c	Assessment: confident appearance
e66013d	Assessment: ambition
e66013e	Assessment: stamina
e66013f	Assessment: empathy
e316210	Assessment mathematical competence
e316220	Assessment German competence
e316251	Individually expected school-leaving qualification

Exemplary data snapshot

ID_t	ID_e	wave	ed1105a	ed1105b	e66013a	e66013b	e66013c
4089432	1023059	2	yes	yes	4	4	3
4090521	1021056	2	yes	yes	2	4	2
4091499	1023284	2	yes	yes	2	3	1
4091533	1023284	2	yes	yes	2	3	4
4097814	1024621	2	yes	yes	2	2	4

This file contains data from the individual assessments of the participating "NEPS students" by their class teachers. The data collection took part in the second survey wave. The topics covered range from behavioral traits such as diligence, ambition, and hyperactivity (`ed66013*`) to mathematical and German competence (`e316210/20`), as well as to the expected school-leaving qualification (`e316251`).

It is important to note that, on the one hand, several students or target children were assessed by the same class teacher. This means that the identifier variale `ID_e` is not unique in the dataset. On the other hand, some students were assessed multiple time by different teachers (up to five individual assessments per target child). This means that identifier variable `ID_t` is also not unique in the dataset.

Stata 14: Working with pTargetAssessments

```
** open the data
use "${datapath}/SC8_pTargetAssessments_D_${version}.dta", clear

** notice that for some students there are multiple assessments by different teachers
duplicates report ID_t wave

** for merging illustrations, we only keep one assessment for each students
bysort ID_t wave: keep if _n==1

tempfile assessments
save `assessments'

** open CohortProfile data
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear

** merge assessments
merge 1:1 ID_t wave using `assessments', keep(master match)
```

4.5.13 TargetMethods

« go back to overview

Description

Paradata from the targets CASI interview

File structure

Long format: 1 row = 1 target in 1 wave

ID variables needed to identify a single row

ID_t wave

Other ID variables useful for linkage

none

Number of variables / number of rows in file

13 / 12,166

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave	Wave
tx80201	Interview: survey mode (start)
tx80202	Interview: Survey mode (realized case)
tx80221	Interview: evaluable data set?
tx80400	Willingness: panel participation

Exemplary data snapshot

ID_t	wave	tx80201	tx80221	tx80400
4088119	1	CASI / TBT	does apply	yes
4088560	1	CASI / TBT	does apply	yes
4089432	2	CASI / TBT	does apply	yes
4089433	2	CASI / TBT	does apply	yes
4091245	2	CASI / TBT	does apply	yes

This dataset offers information on the data collection during the CASI interview with the target children in the school context. These include, for example, the realized survey mode (`tx80202`), the differentiated response code (`px80207`), or the given incentive (`tx80210`). Other information refer to the sampling frame such as federal state (`tx80101_R`) or municipality size (`tx80102/tx80103`) as well as to the tracking list such as transition recommendation to secondary school (`tx82001`) or special educational needs (`tx820*_R`).

Please note that this file contains all participating target children, whether an interview was realized or not. Thus, `TargetMethods` includes more cases than the data file `pTarget`. It is also important to know that the target ID (`ID_t`) is persistent across waves.

Stata 15: Working with TargetMethods

```
** open the data file
use "${datapath}/SC8_pTarget_D_${version}.dta", clear

** merge variable tx80210 "Interview: incentive (euro)" to data
merge 1:1 ID_t wave using "${datapath}/SC8_TargetMethods_D_${version}.dta", ///
        nogen keep(master match) keepusing(tx80210)

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

** view incentive vs. household size
tab t741002 tx80210
```

4.5.14 Weights

« go back to overview

Description

Sample weights for various occasions

File structure

Wide format: 1 row = 1 target

ID variables needed to identify a single row

ID_t

Other ID variables useful for linkage

ID_i

Number of variables / number of rows in file

17 / 6,141

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
ID_i	Identifier for the institution
sample	Sample that target person belongs to
stratum_exp	Explicit stratum (school type according to sampling frame)
w_i	Final weight for institution, calibrated
w_t1	Cross-sectional weight for targets participating in wave 1, calibrated

Exemplary data snapshot

ID_t	ID_i	sample	stratum_exp	w_i
4090043	1004978	K5	3	30.41671
4089562	1005004	K5	1	44.41052
4090054	1005228	K5	4	17.61941
4089673	1005180	K5	1	31.84000
4087329	1005077	K5	1	17.54435

This dataset includes weighting variables (`w_`) as well as sample stratification characteristics (`stratum_`). Given the complex structure of the sample, there is no final recommendation at hand concerning the best use of the design weights (both for the schools as institutions and for the target children) or the cross-sectional weights (for participation per survey wave). More detailed information about weights estimation can be found in the wave-specific reports on “Samples, Weights and Nonresponse” as part of the data documentation (see Section 1.2).

Stata 16: Working with Weights

```
** open CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear

** merge Weights data
merge 1:1 ID_t using "${datapath}/SC8_Weights_D_${version}.dta", nogen keep(master
  match) keepusing(w_t_cal_std)

** you now have the variable w_t_cal_std available for further usage
```

4.5.15 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

105 / 5,738

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave_w1	Row contains data from wave 1 (2022/2023)
wave_w2	Row contains data from wave 2
mbg5_pv1b	Mathematical competence (education trend): plausible value 1 (Ed. trend metric)
mbg5_pv2b	Mathematical competence (education trend): plausible value 2 (Ed. trend metric)
rbg5_pv1b	Reading competence (education trend): plausible value 1 (Ed. trend metric)
rbg5_pv2b	Reading competence (education trend): plausible value 2 (Ed. trend metric)

Exemplary data snapshot

ID_t	wave_w1	wave_w2	mbg5_pv1b	mbg5_pv2b	rbg5_pv1b
4089213	yes	yes	532.71217	487.60017	-55.00000
4089345	yes	yes	545.77679	552.71820	-55.00000
4086021	yes	yes	602.28313	458.71348	-55.00000
4088515	yes	yes	545.77679	569.02935	-55.00000
4089517	yes	yes	441.96818	379.84907	-55.00000

Plausible Values (PV) are a way of describing individual competencies at group level. They enable (unbiased) estimates of effects at the group level that are adjusted for measurement errors. 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. Hypothesis tests for the specific question of interest are calculated for each of these values and combined into an overall result (Scharl et al., 2020).

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

Stata 17: 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!
```

4.5.16 xTargetCompetencies

« go back to overview

Description

Test data of target children at regular schools

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

657 / 5,742

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave_w1	Row contains data from wave 1 (2022/2023)
mag5d041_sc8g5_c	Mathematical competence: item 1
mag5_sc1	Mathematical competence: WLE (corrected)
reg5_sc1	Reading competence: WLE (corrected)
reg5_sc2	Reading competence: standard error of WLE (corrected)

Exemplary data snapshot

ID_t	wave_w1	mag5d041_sc8g5_c	mag5_sc1	reg5_sc1	reg5_sc2
4086404	yes	1	1.51	1.06	0.61
4087096	yes	1	0.45	1.28	0.57
4089627	yes	1	2.37	1.31	0.82
4090101	yes	1	0.86	1.41	0.65
4090837	yes	1	1.23	0.87	0.68

The file `xTargetCompetencies` contains data from competence assessments conducted with the target children in **regular schools**. Scored item variables as well as scale variables are available in a cross-sectional format. Table 2 shows which competencies have been tested and when those testings have been conducted. In addition to the naming conventions for competence variables (see Section 3.2.2), the variables `wave_w*` are provided to select rows only containing competence data from a specific wave.

Stata 18: Working with xTargetCompetencies

```
** open datafile
use "${datapath}/SC8_xTargetCompetencies_D_${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 take part in the testing
drop if wave_w1==0

** and reduce the dataset to the relevant variables
keep ID_t wave mag5_sc1 mag5_sc2

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

** and merge this to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge 1:1 ID_t wave using `tmp', nogen
```

4.5.17 xTargetSpecialNeedsCompetencies

« go back to overview

Description

Test data of target children at special schools

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

415 / 229

Contains data from waves

1 2

Exemplary variables

ID_t	Identifier for target person
wave_w1	Row contains data from wave 1 (2022/2023)
wave_w2	Row contains data from wave 2 (2023/2024)
nig50101_sc8g5_c	Non-verbal reasoning (figure classification): item 1
nig5_sc1	Non-verbal reasoning (figure classification): WLE (corrected)
reg50110_sc8g5_c	Reading competence: item 1
reg5_sc1	Reading competence: WLE (corrected)
reg5_sc2	Reading competence: standard error of WLE (corrected)

Exemplary data snapshot

ID_t	wave_w1	wave_w2	nig50101_sc8g5_c	nig5_sc1	reg50110_sc8g5_c	reg5_sc1
4091598	yes	yes	1	1.0663	1	0.22
4091472	yes	yes	1	1.4283	1	2.00
4091448	yes	yes	1	-0.8215	1	1.05
4091483	yes	yes	1	0.0117	1	-3.35
4091517	yes	yes	1	1.0663	1	0.47

The file `xTargetSpecialNeedsCompetencies` contains data from competence assessments conducted with the target children in **special education schools**. Scored item variables as well as scale variables are available in a cross-sectional format. Table 2 shows which competencies have been tested and when those testings have been conducted. In addition to the naming conventions for competence variables (see Section 3.2.2), the variables `wave_w*` are provided to select rows only containing competence data from a specific wave.

Stata 19: Working with xTargetSpecialNeedsCompetencies

```
** open datafile
use "${datapath}/SC8_xTargetSpecialNeedsCompetencies_D_${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 Reading competence, which have been tested in wave 1.
generate wave=1

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

** and reduce the dataset to the relevant variables
keep ID_t wave reg5_sc1 reg5_sc2

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

** and merge this to CohortProfile
use "${datapath}/SC8_CohortProfile_D_${version}.dta", clear
merge 1:1 ID_t wave using `tmp', nogen
```

5 Special Issues

5.1 Modules on education in pParent

The *education modules* for the responding parent and for their partner were redesigned in Starting Cohort 8 to reflect international as well as national educational classifications, in particular ISCED and CASMIN, also for those who obtained their highest educational attainment abroad. For this reason, a distinction was made in the survey instrument between German and foreign educational qualifications. There is one block with questions on a German educational qualification (for the responding parent: `p731831` to `p731834`; for their partner: `p731881` to `p731884`) and one block with questions on a foreign educational qualification (for the responding parent: `p731835` to `p731846`; for their partner: `p731885` to `p731896`). For the "German" block, the usual questions are asked to differentiate between school-leaving qualifications and vocational training or higher educational qualifications (see for example the demographic standards). The "foreign" block, however, avoids the use of answer categories that are typical for the German school and educational system as well as for the above-mentioned distinction. It is rather based on generally understandable questions that cover the highest educational qualification in other educational systems as precise as possible (see Schneider et al., 2023).

The first question of the education modules serves to identify whether the highest qualification achieved was a German qualification or a foreign qualification (for responding parent: `p731830`; for his or her partner: `p731880`). This initial question separates the two blocks. Nevertheless, it should be possible for the respondents to switch from one block to the other, as educational careers are diverse and not necessarily limited to one country only.

Unfortunately, a technical problem occurred in programming the instrument for the first survey wave. This prevented the intended "backward jump" from a subsequent block to a previous block. In order to solve this problem, the questions of the previous "German" block were copied and pasted again after the subsequent "foreign" block to enable the aforementioned switch. Answers to these "German" questions inserted at the end were – if available – automatically assigned to the original "German" questions at the beginning. Therefore, the copies of these questions available in the original field instrument are not included as additional variables in the Scientific Use File.

5.2 School history in pTarget

In the first survey wave in grade 5, previous school careers were recorded using an episodic approach in cases where the school currently attended was **not** the first school after the transition to lower secondary level and/or the grade had already been repeated. The episodic survey began when it was reported that the target child had already attended grade 5 before the point of measurement (td00103).

Although this type of episodic recording had been tested in advance, it could not be implemented within the given time frame due to primarily technical issues during the programming of the survey instrument. Also, content-related problems – e. g., the wording of question td00104, which was supposed to refer to the time in the 5th grade and not to the time after that – could not be corrected in due time, so that the instrument was used in its existing form at the beginning of the fieldwork.

In the course of checking the data from the first survey wave before the Scientific Use File was published, the data quality problems proved to be more serious than assumed at the beginning of the fieldwork. For this reason, the strategy used for the programming of the third survey wave has been applied **ex post** to the data of the first survey wave during the process of data editing.[14] This means that only the data from the first episode, i.e. the first school attended in grade 5 – along with some additional information – is available for use. This ensures comparability with the data collected in the third survey wave. The available information, which will be included in the Scientific Use File both for the first wave and later for the third wave, is as follows:

- td00104* "School history: comparing current school A"
- td00106 "School history: school attendance in Germany A"
- td00108* "School history: country of school"
- td00112* "Federal state/Municipality of school"
- td00115 "School history: school authority"
- td00116 "School history: pedagogical concept of the school"
- td00117/td00118* "School history: type of school" depending on the federal state specified (including open text answers)
- td00120/td00123* "School history: first school branch other school" (including open text answers for schools with school tracks)

14 Similar problems or challenges that could not be solved in the available time frame also arose when programming the third wave of Starting Cohort 8 for the newly surveyed students in grade 7. A comparable episodic recording was planned for these persons. During the programming process, it was therefore decided to replace the episodic measurement with a simplified cross-sectional module with selected questions.

For the first wave – in contrast to the third wave – the variables relating to the same school, namely td00119 "School history: several school branches" and td00121 "School history: first school branch current school", are also published in the Scientific Use File.

Furthermore, the validity of the answers to the incorrectly formulated initial question has been checked by comparing and combining these answers with the responses of the students to the subsequent episode questions.[15] The result of this check is reflected in the recoded variable td00104_g1 in the Scientific Use File. If the provided information for the episode were contradictory, this variable has been coded with the value 3 "contradictory statements" and the subsequent variables were assigned the missing code –52 "implausible value removed".

Finally and according to the theoretical intention of the instrument, the variable td00106 "School history: school attendance in Germany" has been recoded with the value 1 "yes" if in variable td00104_g1 the answer "I was in the same school as now" was given. For these cases, the variable td00108 "School history: country of school" was additionally coded to "Germany".

5.3 Individually followed students

Students who leave the originally sampled "NEPS schools" in Starting Cohort 8 are followed or tracked individually as part of the longitudinal study design. The same applies to students enrolled at schools that withdrew from participating in the NEPS study. Beginning in survey wave 2 (sixth grade), these target children are first contacted directly via telephone by the survey agency. After completion of a computer-assisted telephone interview (CATI), the students are invited to answer a self-administered questionnaire, which is provided either in paper form (PAPI) or as a computer-assisted web interview via secure online survey platform (CAWI) – depending on the participants' preference. In cases where the telephone survey part cannot be realized, the students are asked to complete an extended version of the paper or online questionnaire. This extended version includes some more items to also collect the information for this group that would otherwise be gathered in the CATI survey.

In addition to the regular survey instruments, the instruments administrated to the individually followed or tracked students contain specific questions about their current educational placement (e.g., type of school and enrollment status). In contrast to the survey in the school setting, the survey among individually followed or tracked students does not include any questions about aspects of the class context.

15 The episode information used for the consistency check are based on continuation of the school episode at the school or school track, reason for change to another school or to current school and school attendance in Germany.

5.4 Survey of school staff

In the NEPS Starting Cohort 8 – unlike in the previous Starting Cohort 3 – the entire teaching staff, including school administrators, at each participating school was invited to participate in the study via online questionnaire at the start of the panel (first wave). As a result, the corresponding datasets also contain information from teachers or about classes that are not linked to any students (target children) in the study. Thus, the `pEducator` dataset comprises information on teaching staff members who cannot be linked to the survey and competence data of one or more students, since at the time of the survey, the respondent did not hold any position as a class teacher, German teacher, or mathematics teacher for any of the "NEPS students". Currently, such teachers can only be identified through a *merge* operation with the `LinkTargetTeacher` dataset, where they remain as unlinkable cases. Due to the survey design, the function-specific `pCourse*` datasets also contain information about classes in which not a single student participated in the NEPS survey. These cases can be detected using the additionally generated variable `ex20103` ("Participation students from class").

In the second wave, only class teachers were surveyed (see also Section 5.5). Furthermore, questionnaires were provided only to those school administrations that had not already participated in the NEPS survey during wave 1. The questionnaire was a shortened version of the original questionnaire from the first wave. In this short version, all information was collected retrospectively for the period of the first wave – that is, the questionnaire included instructions to refer the responses to the school year 2022/23. The invitation to participate in this survey was sent to the entire school administration team at each of the respective schools. In theory, this created the possibility of multiple respondents per school for wave 2 as well. In practice, however, this did not occur.

5.5 Individual student assessments

In the second survey wave, the participating class teachers assessed various characteristics and the expected school-leaving qualification for each of their "NEPS students". Class teachers who also taught German and/or mathematics were additionally asked, as part of these individual assessments, about each student's competences in those subjects. The corresponding data are stored in the `pTargetAssessments` dataset (see Section 4.5.12). For the fourth survey wave, another round of individual assessments of students by their teachers is planned.

Since some schools have teams of class teachers, the Scientific Use File contains multiple assessments for some target children – provided by different persons.

It should also be noted that class teachers who participated in the NEPS survey for the first time during the second wave answered a few additional questions about themselves. This information is integrated into the `pEducator` dataset (see Section 4.5.8).

5.6 Survey of parents

In principle, the survey design aims to always interview the same parent of a target child repeatedly – whenever possible. However, the participating parent may change across survey waves. The reasons for this are manifold and range from parents' availability to their motivation and changes in primary responsibility for the child's care, to separation and death or the loss of legal custody. However, even if the parent being interviewed changes, it is ensured that the information provided is consistently linked to the correct person over time via a persistent identifier (`ID_p`).

Due to changes of the so-called "anchor person", the sociodemographic details of the parents may be spread across different survey waves. At each person's first NEPS parent interview, their sociodemographic information and certain characteristics of their respective partner are comprehensively collected. When a parent is surveyed for the first time – meaning no parent interview has been conducted with that person previously – sociodemographic information related to the target child, as well as information on the child's school history, is also collected. This initial survey of a parent can take place at any time during the course of the panel study. As long as the parents have not withdrawn their consent to participate, they will continue to be invited to every NEPS survey, even if an interview could not yet be realized.

In addition, in each survey wave there is an update of certain information provided by the parents participating in the panel survey. In other words, sociodemographic data may be available even beyond the recruitment time point (first parent interview).

5.7 Recoding of system missing values

Due to the use of new survey software, missing values that were originally stored as "system missing" (.) have been recoded during the subsequent data editing process for the Scientific Use File of NEPS Starting Cohort 8. The following recoding rules apply to all datasets:

- The value -99 ("filtered") is assigned when a respondent was not asked a particular question. There are two possible reasons for this: Either the question was intentionally filtered out, or the respondent did not see the question for other reasons. One such reason is the abortion of the survey at a point preceding the variable – without the processing routine for identifying abortions having correctly detected this (otherwise, the value would be recoded to -91). Another possible reason is that the respondent intentionally skipped a question. In this case, the system should actually assign the missing code -90 ("unspecific missing"). However, the programming for this did not function perfectly in the first two survey waves.
- The value -90 ("unspecific missing") replaces a system missing value whenever a respondent sees the question but has not been given any explicit options to refuse to answer – and the respondent has deliberately chosen not to answer this question.

- The value -91 (“survey aborted”) is subsequently recoded in the data editing process if a detection routine identifies a previous survey abortion. Essentially, this routine works by starting at the end of a survey instrument and, in reverse order, assigning the value -91 to all variables with missing values – until the first valid value appears in the dataset. A distinctive feature of the new survey software is that, for some item batteries the value 0 is automatically given by the system if a respondent has already terminated the survey. Since the value 0 can also represent a meaningfully interpretable response for certain variables, this value is overwritten with the missing code -91 only in cases where all variables in the relevant item battery have the value 0.

If, during the preparation of data for the Scientific Use File, it is discovered that a data row consists entirely of missing values – that is, contains no information provided by the respondent – then this data row is deleted. Under certain circumstances, such “empty” data rows result from respondents logging into an online questionnaire multiple times without actually entering any information. Therefore, this cleaning does not affect any informational value of the data.

A References

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. (2026). *Data Manual of NEPS Starting Cohort 8 – Grade 5 (2022), Education for the World of Tomorrow, Scientific Use File Version 2.0.0*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gnambs, T. (2025a). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 for Grade 5* (NEPS Survey Paper No. 117). Leibniz Institute for Educational Trajectories. https://doi.org/https://doi.org/10.5157/NEPS:SP117:1.0

Gnambs, T. (2025b). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 for Grade 5 in Special Schools* (NEPS Survey Paper No. 118). Leibniz Institute for Educational Trajectories. https://doi.org/https://doi.org/10.5157/NEPS:SP118:1.0

Gnambs, T. (2025c). *NEPS Technical Report for Verbal and Nonverbal Reasoning: Scaling Results of Starting Cohort 8 for Grade 5 in Special Schools* (NEPS Survey Paper No. 119). Leibniz Institute for Educational Trajectories. https://doi.org/https://doi.org/10.5157/NEPS:SP119:1.0

Hellrung, M., Hillen, P., Hugk, N., Meyer-Everdt, M., Sievers, U., & Tusch, S. (2025). *Feld- und Methodenbericht der IEA Hamburg zur NEPS-Teilstudie A104*. Hamburg, Germany: IEA.

Konrad, A., Würbach, A., & Aßmann, C. (2025). *Samples, Weights and Nonresponse: NEPS Starting Cohort 8 – Grade 5 (2022). Education for the World of Tomorrow. Wave 1*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

NEPS Network. (2026-a). *National Educational Panel Study, Scientific Use File of Starting Cohort 8 – Grade 5 (2022)*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC8:2.0.0.

NEPS Network. (2026-b). *Starting Cohort 8 – Grade 5 (2022), Wave 2, Questionnaires (SUF Version 2.0.0)*. Bamberg, Germany, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pelz, S. (2025). *NEPS Technical Report: Implementation of the ISCED-2011, CASMIN and Years of Education Classification Schemes in SUF Starting Cohort 8*. 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.

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

Schneider, S., Chincarini, E., Liebau, E., Ortmanns, V., Pagel, L., & Schönmoser, C. (2023). *Die Messung von Bildung bei Migrantinnen und Migranten in Umfragen*. Mannheim, GESIS – Leibniz- Institut für Sozialwissenschaften (SDM – Survey Guidelines). https://doi.org/DOI:10.15465/gesis-sg_040

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

B Appendix

B.1 Release notes

Below you can find the *Release Notes* for the current Scientific Use File. They contain information on relevant data edition issues compared to the previous version of the Scientific Use File as well as information on data-related specifics and known problems at the time of the data publication. The *Release Notes* can also be downloaded from the documentation website of Starting Cohort 8 – as a text file with the complete history of data edition information on all Scientific Use File versions so far (see Section 1.2).

```
************************************************
**
** NEPS STARTING COHORT 6 - RELEASE NOTES
** Changes and Updates for SUF Version 17.0.0
** (doi:10.5157/NEPS:SC6:17.0.0)
**
************************************************


General:
        - data from wave 17 have been newly integrated into the Scientific Use File
        - meta Information/labels were reviewed for all variables and updated as needed

CohortProfile & Methods:
        - information regarding the linkage of SC6 NEPS survey data with administrative
          data from the Institute for Employment Research has been updated and now
          refers to the data package "NEPS-SC1-ADIAB 7524 "v1 provided by the 'IABs
          Research Data Center (Methods: tx80130, tx80132 | CohortProfile: tx80533)

xPlausibleValues:
        - plausible values were recalculated for all Waves
        - plausible values for the competence measures from wave 14 were generated and
          made available  for the first time

xTargetCompetencies:
        - data from the competence measurements of wave 17 are currently still being
          processed and scaled; this data material will be published immediately
          thereafter as part of an update to the present Scientific Use File
```