NEPS
National Educational Panel Study

Research Data

Starting Cohort 4: Grade 9 (SC4)
SUF Version 1.1.0
Data Manual

Jan Skopek, Sebastian Pink, Daniel Bela

SPONSORED BY THE

Federal Ministry
of Education
and Research

Research Data Papers
at the NEPS Data Center, Bamberg

The NEPS Research Data Paper series presents documentation resources prepared to support the work with data from the National Educational Panel Study (NEPS).

Full citation of this document:
Jan Skopek, Sebastian Pink, and Daniel Bela (2013). *Starting Cohort 4: 9th Grade (SC4). SUF Version 1.1.0. Data Manual.* NEPS Research Data Paper. National Educational Panel Study (NEPS), University of Bamberg

This release of scientific use data from Starting Cohort 4 – "School and Vocational Training" was prepared by the staff of the NEPS Data Center. It represents a major collaborative effort. The contribution of the following staff members of the NEPS is gratefully acknowledged:

Data preparating and editing
Dietmar Angerer (filtering syntax development)
Daniel Bela (file integration, data preparation, integration of metadata)
Tobias Koberg (anonymisation, regional data, translation)
Manuel Munz (coding and classification)
Sebastian Pink (data preparation)
Jan Skopek (file design and data preparation)
Knut Wenzig (management and editing of metadata, documentation)
NEPS pillar 1a (scoring and scaling of competence data)

National Educational Panel Study (NEPS)
Data Center
Wilhelmsplatz 3
96047 Bamberg, Germany

E-mail: userservice.neps@uni-bamberg.de
Web: https://www.neps-data.de/en-us/datacenter
Phone: +49 951 8633511

This document has been typeset with XeLaTeX 0.9998 using the free font Linux Libertine from the Libertine Open Fonts Project.

Contents

List of Tables

List of Figures

List of syntax examples

1 Introduction

1.1 About this manual

This manual is intended to assist your work with the scientific use data of the NEPS Starting Cohort 4—School and Vocational Training (NEPS SC4, release 1.1.0). We aim at providing a detailed guide of how to use these data for your research. Therefore, our focus is on practical aspects of data usage such as key aspects of the survey and sampling design, the dataset structure, key variables, and syntax examples.

This manual is not a complete and exhaustive documentation resource. Please consult our website[1] for background information on the studies, survey instruments, structured documentation, and many more resources.

We aim at keeping this manual as short and simple as possible. At several places, we reference supplementary documents presenting additional information that we consider essential for working with our data:

Table 1: Documentation supplements for Starting Cohort 4

Description	Language
Codebook (Wenzig 2012c, Wenzig 2012b)	DE,EN
Survey instruments package	DE,EN
Technical report on weighting (Aßmann, Steinhauer, and Zinn 2012)	EN
Technical report on competence scaling (Pohl and Carstensen 2012)	EN
Technical report on anonymization (Koberg 2012a)	EN
Technical report on regional data (Koberg 2012b)	EN
Technical report on data edition (forthcoming)	EN
Field work report infas (Aust, Hess, and Prussog-Wagner 2013, forthcoming(a))	DE
Field work report IAE-DPC (forthcoming)	DE
Syntax package (e.g., cleaning syntax for PAPI data)	Stata

You can download these documents by navigating through our website to:

→ www.neps-data.de > Data Center > Research Data > Starting Cohort 9th grade

We welcome feedback from our users that will help us improve the quality of this manual and our data for future releases. Please report any feedback to:

- userservice.neps@uni-bamberg.de

1 https://www.neps-data.de/en-us/datacenter

1.2 Release history

- *March 22, 2013*: Updating release version 1.1.0. Release Notes: `https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-1-0/10_5157_NEPS_SC4_1_1_0.txt`
- *October 26, 2012*: Data from Starting Cohort 4 are available in release version 1.0.0.

1.3 Obtaining the data

There are three simple steps to obtain the data of this release:

- Sign the data use contract and mail it to us:

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

- After approval, sign in as a registered NEPS user at the login at `https://www.neps-data.de`
- Access the data via one of our three access modes (see below)

Depending on which access mode(s) you choose, you will find all further instructions required to access the data on our website.

1.4 Three modes of data access

Once you have signed a data use contract with the NEPS, we provide to you three modes of access:

- Download from our website (after login),
- RemoteNEPS (browser-based remote access),
- and on-site access.

These three solutions are designed to support the full range of users' interests and maximize data utility while complying with strict standards of confidentiality protection. Access via *RemoteNEPS* works with *biometrical authentification* and requires *at least one participation in the user training courses* provided by the NEPS Data Center.

Sensitive data

Each access mode corresponds to a specific level of data sensitivity. Files that are offered for download include data with the highest level of anynomization. These data are available to registered users from the web portal via a secure connection. Files offered via RemoteNEPS contain more sensitive data within a controlled environment. The analysis of information in high resolution (e. g., fine-grained regional information) is only provided on-site in Bamberg

where these data are available within a secure site. For details on the access modes, see our website at

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

This concept of data dissemination translates into an "onion-shaped" model of datasets: The most sensitive data ("on-site") that include information in high resolution and minor anonymization represent the outer layer, with "remote access" and "download" levels being subsets of these data. That is, any data contained within a less sensitive level is also included in the higher level(s).

An overview on which types of data are offered at each of these levels as well as detailed information on how the "on-site", "remote access" and "download" versions of the data were generated can be found in technical report on anonymization (Koberg 2012a, see section 1.1).

File Format

All files are available in Stata and SPSS format with bilingual variable labels and value labels (German and English). Data stored in Stata format contain both languages within one file (see section 9 for a tutorial on how to switch label language in Stata). SPSS files are provided by language separately.

1.5 Respect rules of data usage!

When working with the NEPS data be aware of the data usage rules you have signed in the NEPS data contract! In particular, in the context of this NEPS data release you are not allowed to publish any analyses that *aim for* or *allow* a *direct comparison of the German federal states (Bundesländer).* Any forms of "rankings" of German Bundesländer using the NEPS data is strongly prohibited. Also *singling out particular Bundesländer* for analyses is prohibited! Note, that all variables identifying Bundesländer have been removed from the data files of this release. Read more in section 8.1.

1.6 Publications with NEPS data

If you publish with NEPS data, it is mandatory to quote the following reference:

> Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, eds. (2011). *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).*

In addition, publications using data from this release must include the following acknowledgement:

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 4 – 9th grade (School and Vocational Training), doi:10.5157/NEPS:SC4:1.1.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

A digital object identifier (DOI) uniquely identifies each release of NEPS data (cf. Wenzig 2012a). The DOI of this release redirects to a landing page providing basic information on the data:

SC4 Version 1.1.0 `doi:10.5157/NEPS:SC4:1.1.0`
SC4 Version 1.0.0 `doi:10.5157/NEPS:SC4:1.0.0`

2 General Conventions

2.1 File names

The naming of data files included in this release follow a number of conventions which are summarized in table 2 on the following page.

To give an example, the physical file `SC4_xTarget_D_1-1-0.dta` refers to the *download-version* for the data file *xTarget* of *starting cohort 4* of data release *1.1.0*.

2.2 Variable names

The variable naming conventions are aimed at ensuring the consistency of variable names across panel waves. They reflect the panel structure of the NEPS data and allow users to conveniently identify variables across waves. General conventions for variable names are presented in section 2.2.1. Variables corresponding to test items (competence assessments) follow a separate nomenclature that is optimized for working with competence data. Rules for naming competence variables are introduced in section 2.2.2.

Table 2: Naming conventions of file names

Element	Definition
SC[1-6]	**Indicator of starting cohort** 1 = Infants 2 = Kindergarten 3 = 5th grade students 4 = 9th grade students 5 = First-year undergraduate students 6 = Adults
[filename]	**Filename conventions** *Prefix*: x = cross-sectional file; sp = spell file; p = panel file *Keyword/mnemonic*: A keyword or mnemonic indicates the content of the corresponding file (e.g., data file xTarget contains cross-sectional data from the target questionnaire; spSchool contains spell data from school history spells) Filenames of generated datasets do not have a prefix and always start with a capital letter (e.g., CohortProfile, Biography)
[D,R,O]	**Confidentiality Level** D = Download version R = Remote access version O = Onsite version
[#]-[#]-[#](_beta)	**Version** *First digit*: denotes the main release number; the main release number is incremented with every further wave release of a starting cohort; however, the first digit does not imply the number of waves included in the release; e.g., in starting cohort 4, the main release number 1 comprises already two wave data (first wave fall 2010 and second wave spring 2011). *Second digit*: indicates major updates; major updates affect the data structure (e.g., release of imputed datasets); updating your syntax files may be necessary. *Third digit*: indicates minor updates; minor updates affect the content of cells but not the data structure; updating your syntax files is not necessary. _beta: this suffix indicates a preliminary release which allows users to test the data in advance of the main release. The beta version is no longer available after the main release.

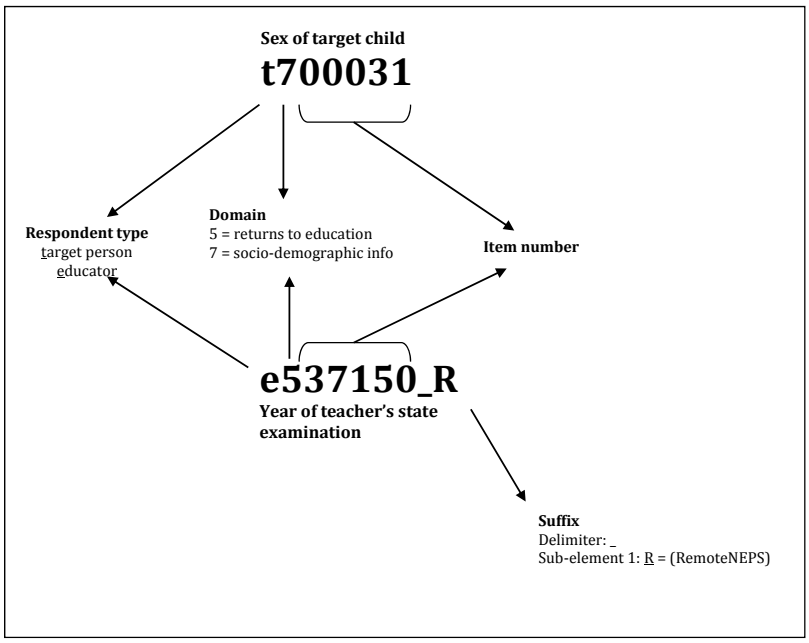


Figure 1: Example for variable naming

2.2.1 General naming conventions

A variable name consists of up to four elements: the respondent type, the domain of information, an item number, and an optional suffix providing further information e. g. whether the variable is generated and/or only accessible via RemoteNEPS. The principles of the naming conventions are illustrated by figure 1. More detailed information is given in table 3.

Table 3: Naming conventions of variable names

Digit	Description
1	Indicates to which **respondent type** the variable refers; in starting cohort 4, this character can be `t` (target person, `p` (one parent of target person), `e` (educator) and `h` (information about the school/kindergarten given by the principal). Sometimes, for the sake of usability names of variables relating to the target begin also with a `t` even if the target was not the actual respondent. For example this is usually true for generated variables and variables containing para data (e.g. list data from the schools/kindergartens).

(...)

Table 3: (continued)

Digit	Description
2	**Topic/domain** (according to the theoretically coordinated dimensions of the NEPS): 1 = competence development (pillar 1) 2 = learning environments (pillar 2) 3 = educational decisions (pillar 3) 4 = migration background (pillar 4) 5 = returns to education (pillar 5) 6 = working group "interest, self-concept and motivation" 7 = socio-demographic information a = from birth to early child care (stage 1) b = Kindergarten to elementary school (stage 2) c = from elementary school to lower secondary school (stage 3) d = from lower to upper secondary school (stage 4) e = from upper secondary school to higher edu./occup. training/labor market (stage 5) f = from vocational training to the labor market (stage 6) g = from higher education to the labor market (stage 7) h = adult education and lifelong learning (stage 8) s = basic program x = generated variables
3–7	**Item number**: The item number typically consists of four numeric characters plus one alphanumeric character.
8–11	**Suffix** (optional): Suffixes are separated from the previous characters by an underscore. There are three types of suffixes: **Suffixes for generated variables** Generated variables are indicated by the suffix _g# (_g1, _g2, etc.). In most cases, the running number after _g is a simple enumerator. However, there are two types of generated variables that assign meanings to these running numbers: regional and occupational variables. *Regional codes based on the Nomenclature of Territorial Units for Statistics (NUTS)* g1 = NUTS level 1 (federal state/Bundesland) g2 = NUTS level 2 (government region/Regierungsbezirk) g3 = NUTS level 3 (district/Kreis)

(...)

Table 3: (continued)

Digit	Description
	Occupational/prestige codes
	`g1` = KldB 1988 (German Classification of Occupations 1988)
	`g2` = KldB 2010 (German Classification of Occupations 2010)
	`g3` = ISCO-88 (Internat. Standard Classification of Occupations 1988)
	`g4` = ISCO-08 (Internat. Standard Classification of Occupations 2008)
	`g5` = ISEI-88 (Internat. 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)
	`g10` = DKZ 2010 (Documentary Code Number 2010)
	`g11` = DKZ 1988 (Documentary Code Number 1988)
	`g12` = Coding scheme
	`g13` = KKZ (Course code/Kurskennziffer)
	`g14` = ISEI-08 (Internat. Socio-Economic Index of Occupational Status 2008)
	`g15` = CAMSIS (Social Interaction and Stratification Scale)
	`g16` = SIOPS-08 (Standard International Occupational Prestige Scale 2008)

As scales are generated by a set of other variables, they are also indicated by the above mentioned nomenclature. For the sake of completeness and clarity, it has to be stated that scales are named according to the first variable of the sequence they were generated from. Their running numbers are in so far meaningful as they count up if and only if the first variable of two scales had been identical.

Wide-format suffix

Wide-format variables stored are indicated by the suffix `_w#` (e.g., `_w1`, `_w2`, etc.). Note that the wide-suffix not necessarily implies a wave logic. For instance, the presence of a set of variables `a_w1`, `a_w2`, ..., `a_w10` means that there are up to 10 values for the variable `a` (e.g. the item corresponding to variable `a` was measured repeatedly in a questionnaire loop) relating to a row entity (e.g. a person or a school episode). Of course, there are cases where suffix `_w#` directly relates to wave-specific values of the underlying variable.

Confidentiality suffix

This suffix pertains to all variables that were anonymized (see section 1.4 on page 2). The suffix indicates a variable's degree of anonymization. This suffix may either stand alone (e.g., country of birth: `t405010_R`) or be combined with other suffixes (e.g., district of place of birth: `t700101_g3R`).

(...)

Table 3: (continued)

Digit	Description
	O on site; data on this variable are only available on site
	R remote access; data on this variable are available on site or via RemoteNEPS
	D download; data on this variable are available via all three modes of access

2.2.2 Special conventions for variables in test data

Naming of variables corresponding to test items (usually found in competence data files) follow an alternative nomenclature. Variable names consist of three parts and additional suffixes. The first part defines the test instrument (two characters, e. g. `vo` for vocabulary), the second part defines the target group (two characters, e. g. `k1` for children in Kindergarten in the first wave, i. e., 2010), and the third part defines the item number.

Table 4 on page 11 gives an overview to the logic of parts. The first two characters identify competence domains. An overview of the different competence domains is given in the first part of table 4. The target group indicates the cohort or testing wave in which the item was first used. The different target groups are listed in the second part. In some tests, (e. g., mathematic competence tests) items are implemented in different testing waves. In these cases, the variable name contains the target group for which the item was first used. The variable name of the item is then fixed and does not change when the item is used again in later waves or other cohorts (e. g., if the item is first used in grade 5, the second part of the variable name will be `G5`, even when the item is reused in grade 7). Thus, the target group identification in the variable name does not necessarily indicate the cohort or testing wave. However, this labeling rule assures items being used in different studies to have the same variable name. Some competence tests are not designed for specific age groups, but are implemented unmodified in different cohorts and testing waves. The target group of these tests is indicated by `ci` (cohort invariant). The item number is defined differently for different competence domains. For most competence domains they only indicate the different items.

The competence data files contain item variables (responses to the test items) as well as overall competence scores. There are two versions of item variables in competence data: scored items named `[varname]_c` and scored partial credit-items named `[varname]s_c`. For example, `mag9q071_c` is a scored variable measuring that the respective math item—targeted at grade 9 students—was "solved" (value 1) or "not solved" (value 0) by the respondent. Note that the item variable does not necessarily indicate that the students' mathematics skills are measured in grade 5. It could also be that the measurement was done in grade 7 and that an item was used that has already been implemented in grade 5. Additionally to the item responses, overall

measures of the competence score are given. Suffix `_sc[number]` is used for several aggregated scores and the meaning of the suffixed number is fixed as follows: `1`=WLE (Weighted Maximum Likelihood estimates[2]), `2`=standard error of WLEs, `3`=sum, `4`=mean, `5`=difference. For example, variable `grk1_sc3` represents the sum score of the grammar test of children being tested in the first wave (2010) in Kindergarten. Detailed descriptions on how competence scores are estimated can be found in the respective reports for the different competence domains. If there are several aggregated scores (e. g. different sum scores), letters are appended additionally (e. g. `dgg9_sc3a` is of the sum score for perceptual speed, while `dgg9_sc3b` is the sum score for reasoning – both are measures of domain general cognitive functioning).

2 WLEs are estimated in tests that are scaled based on models of item response theory (cf. Pohl and Carstensen 2012).

Table 4: Different parts of names of variables in test data

Part I (2 chars)**: Instrument Meaning**

re	Read	mp	Meta procedural
ma	Math	md	Meta declarative
sc	Science	rs	Reading Speed
ic	ICT	at	Attention
li	Listening	nr	Native Language Russian
vo	Vocabulary	nt	Native language Turkish
or	Orthography	vi	Verbal Intelligence
gr	Grammar	ni	Nonverbal Intelligence
dg	DGCF	fa	FAIR
ef	English Foreign		

Part II (2 chars)**: Target Group Meaning**

n0	Newborn 0	v1	Vocation 1
...	...	...	...
n3	Newborn 3	v3	Vocation 3
k1	KiGa 1	s1	University students 1
k2	KiGa 2	...	...
g1	Grade 1	s5	University students 5
g5	Grade 5	a1	Adults 1: 2009
g9	Grade 9	...	...
ga	Grade 10	a4	Adults 4
gb	Grade 11		
gc	Grade 12		
gd	Grade 13		

ci Cohort invariant (for instruments administered unchanged in all cohorts)

Part III (4 chars)**: Item number**

_

Part IV: Suffix

_c	scored item variable (0=not solved, 1=solved)
_sc1	WLE
_sc2	Standard error of WLE
_sc3	sum
_sc4	mean
_sc5	deviation score (procedural metacognition)
_sc6	proportion of correct items (procedural metacognition)

2.3 Missing values

We provide different missing codes for different situation of missing values. In general, we distinguish between missing codes indicating sorts of item nonresponse, not applicable missings, and edition missings. When working with the NEPS data make sure that you correctly process those codes in your statistical package. Most packages available provide functions for defining missing values. If you use Stata, you can make use of the `nepsmiss` command provided as a part of the nepstools (see section 9). Table 5 provides an overview of missing codes you will encounter in the NEPS data.

Table 5: Overview of missing codes

Code	Missing
Item nonresponse	
−97	refused
−98	don't know
−95	inplausible value
−94	not reached (only applicable for competence tests)
−5/−6/−20,...,−29	item-specific missing with informative value labels
Not applicable	
−54	missing by design (mostly: not included in sample-specific instrument of this wave)
−93	does not apply
.	filtered / system missing (in CATI/CAPI mode)
−90	unknown missing
−99	filtered (in PAPI mode)
Edition missings (recoded into missing)	
−52	implausible value removed
−53	Anonymized
−55	not determinable
−56	Not participated

We distinguish between three types of missing values:

- Item nonresponse occurs if a person did not respond to a question.
  - The most common instances of item nonresponse are refusals (−97) and don't knows (−98).
  - Inplausible values are coded by a −95 missing value.

- For competence data there is a special missing code −94 that indicates that a test item has not been reached, because the target quit the test somewhere before this item.
- further missing codes (−5/−6/−20, ...,−29) pertain to specific nonresponse categories (e. g., variable `p407050_D` indicating citizenship of the target child has a missing code −20 for "stateless").

- Not applicable denotes missing data that occur because the item does not apply to a person. This category comprises two kinds of missings.
  - The first concerns samples: If a question is not included in a sample-specific questionnaire, the code −54 is assigned to all respondents from this sample. This code is used also for the more general case where values of a variable are not available due to design issues.
  - The second concerns individuals: If a question does not apply to a person, it is coded "not applicable" either by the respondent's or the interviewer's remark (−93) or like it is the case for computer-assisted interviews automatically by the survey instrument (. = filtered). In the context of paper-based questionnaires (PAPI mode) the code −99 is set for filtered variables (not by default, but after applying our filtering syntax provided together with the data, see section 4.3).
  - Missings that occur for unkown reasons are coded by −90.
- Edition missings are defined in the process of data editing.
  - Implausible values are recoded into missing (−52).
  - Sensitive information which is only available via RemoteNEPS and/or on site access is anonymized (−53).
  - Coding schemes are used to generate variables (e. g., occupational coding). If the information from the original data is not sufficient to generate a value, we assign the missing code "not determinable" (−55).
  - If a person was not present during the interview, did not fill out a questionnaire although it was administered to her, the concerning variables are assigned the missing code "not participated" (-56). This missing code is special in so far as target persons lacking interview data (e. g. due to illness) usually are not entailed in the corresponding datasets. In the special case of one dataset integrating multiple waves widely this missing code is assigned.

nepsmiss: Recoding missing values in Stata

We offer a Stata ado file on our web portal which automatically recodes all missing values into extended missing values (.a, .b, etc.), and vice versa, while preserving value labels. We generally recommend running `nepsmiss` before any further data preparation. See section 9 for further information on how to install an update the `nepsmiss` command.

3 Surveys and Sampling

3.1 Design of NEPS

The German National Educational Panel Study (NEPS) has been set up to collect data on the acquisition of education, to assess the consequences of education for life courses, and to describe central educational processes and trajectories across the entire life span. The NEPS relies on a thorough life-course perspective as the predominant theoretical frame. Data collection of NEPS is substantially organized by so called stages and pillars. While eight stages cover concepts specific to certain transitions across the life span, longitudinal integration is ensured by five pillars considering the following major theoretical dimensions: competence development, learning environments, social inequalities and educational decisions, educational processes of migrants, and returns to education. As a reference to the NEPS consult Blossfeld, Roßbach, and Maurice (2011). In particular, Blossfeld, Maurice, and Schneider (2011) provide a concise introduction to the major conceptual issues of NEPS.

The NEPS implements a so called *multi-cohort sequence design* that is depicted in figure 2. For obtaining relevant data as quickly as possible, the NEPS starts off with six separate cohorts that are followed up annually. There are two kinds of cohorts. First, there are cohorts that are defined by a specific point in the educational system: *Starting Cohort 2* (Kindergarten children in 2010), *Starting Cohort 3* (fifth graders in 2010), *Starting Cohort 4* (ninth graders in 2010), and *Starting Cohort 5* (first year university students in 2010). Those cohorts are institution-based cluster samples and reflect the major transitions into, within, and out of the general education and vocational training system. Second, there are two age-based cohorts: *Starting Cohort 1* (newborns in 2012) and *Starting Cohort 6* (adults of age 23 to 64 in year 2009). Beyond that, the NEPS also conducts two additional state-specific studies in Thuringia and Baden-Wuerttemberg. Since they are intended to inform on effects of recent school reforms on those states both studies follow a repeated cross-sectional design.

You can find detailed discussions on Starting Cohort 1 at Schlesiger et al. (2011), on Starting Cohort 2 at Berendes et al. (2011), on Starting Cohort 3 and 4 at Frahm et al. (2011), Wagner et al. (2011), Ludwig-Mayerhofer et al. (2011), on Starting Cohort 5 at Aschinger et al. (2011), and on Starting Cohort 6 at Allmendinger et al. (2011).

3.2 Introduction to Starting Cohort 4

This data release comprises data from the first and second wave of the NEPS Starting Cohort 4 (SC4), a sample of ninth graders. One of the major goals of this cohort is capturing the students' transitions and development to upper secondary school and vocational training. For

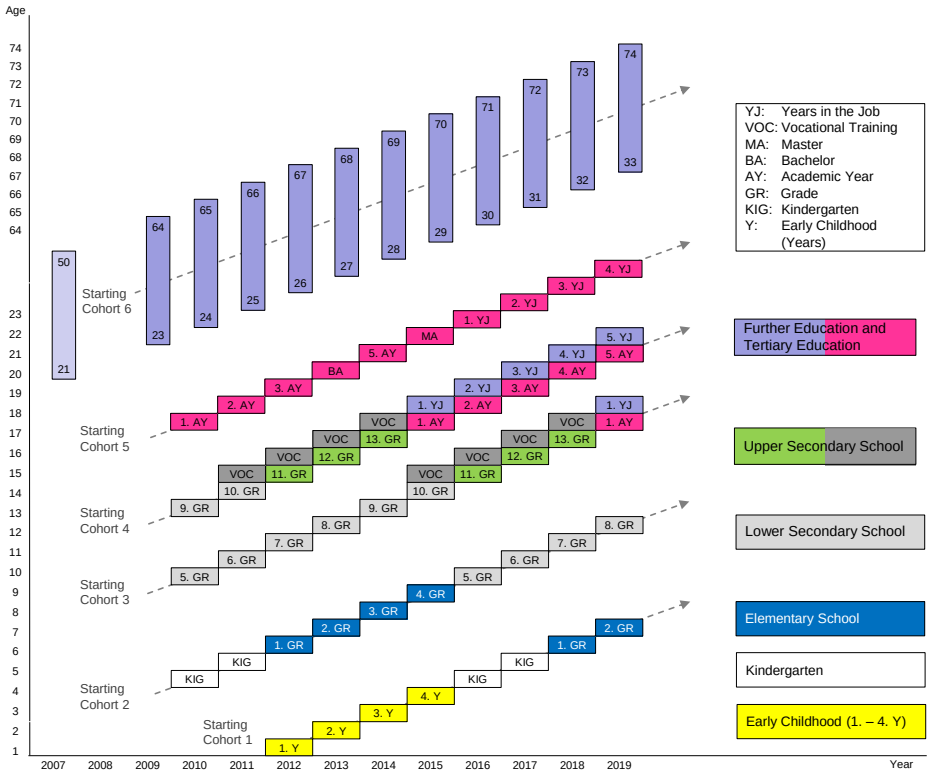


Figure 2: The Multi-cohort Sequence Design of NEPS

an in depth treatment of theoretical perspectives and design issues of the school sample consult Frahm et al. (2011), Wagner et al. (2011), and Ludwig-Mayerhofer et al. (2011).

In general, for starting cohort 4 student samples from two different populations have been drawn: students at regular schools ("Schüler an Regelschulen") and students with special educational needs in the area of learning at special needs schools ("Schüler an Förderschulen"). Both populations were attending 9th grade in German schools in fall 2010. Field work procedures for the special needs sample deviated slightly from those of the regular sample. Hence, when analyzing the data you should always be aware of this fact. Eventually, students were sampled in a multistage approach using an explicit stratification by form of school. Like the cohort of fifth graders, NEPS has initiated a class based sampling design for the ninth graders. That means, entire school classes have been sampled and are followed up over waves.

The surveying and testing in the starting cohort of ninth graders in the school context is also very similar to those of the fifth graders cohort, Starting Cohort 3 (cf. Skopek, Pink, and Bela 2012): Surveying of target persons is organized in the school as group testing and, afterwards, the target persons are asked to fill out a questionnaire. Testing and surveying, in general,

took one school day. Context persons–teachers, principals, and parents–are interviewed by questionnaires very similar to Starting Cohort 3.

Studies

The NEPS Starting Cohort 4 started with two short distance waves. Both waves are part of the current data release. The first wave took place in fall/winter 2010/2011 (data collection from November 22 in 2010 to January 21 in 2011) and the second in spring 2011 (from May 9 in 2011 to July 22 in 2011). Aim of the spring wave was to reassess potential school leavers before leaving school after grade 9. In further waves of Starting Cohort 4, students who drop out of school as well as those who remain within school are followed up. Thus, the starting cohort of the ninth graders is a special cohort because some of the ninth graders have already changed their educational stage from school to vocational education and training (see figure 3). Furthermore, students who are still in the school system, but left the sampled grades, the sampled schools, or their school left the NEPS, are tracked individually with postal questionnaires. In the second wave there were some few students that have been followed up by individual retracking. An overview of the studies and used questionnaires can be found in figure 4 on the next page.

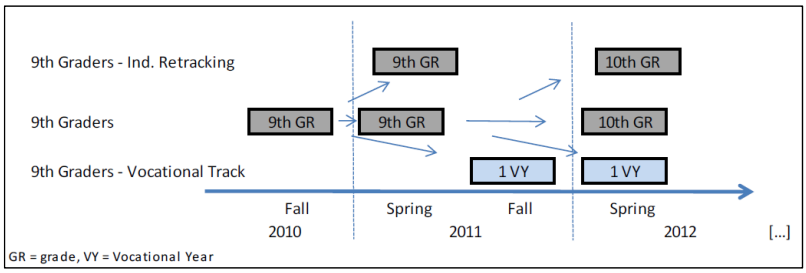


Figure 3: Transitions and Surveys of the Starting Cohort 4

Participation in the NEPS study is voluntary. Parents of children were invited to participate in the study. Note, that contrary to Starting Cohort 2 (SC2), in Starting Cohort 3 (SC3) and 4 (SC4) there was a separate and not a joint consent for parents and children. For example in SC3, parents could agree to the study participation of their child but could refuse their own participation. Thus, the parental consent for their child is the minimum condition for the child being part of the panel sample. In Starting Cohort 4 (9th graders) a separate consent was given by the students.

Data

In total, there are 16,425 ninth graders in the panel sample. At least questionnaire data or test data is available for 15,627 students in first wave. A small number of students could not be surveyed for example due to illness, thus, those are regarded to be temporary dropouts. In the spring wave (second wave) data from 15207 students is available. Some students were not surveyed in the class, but were individually tracked. A few students already switched to

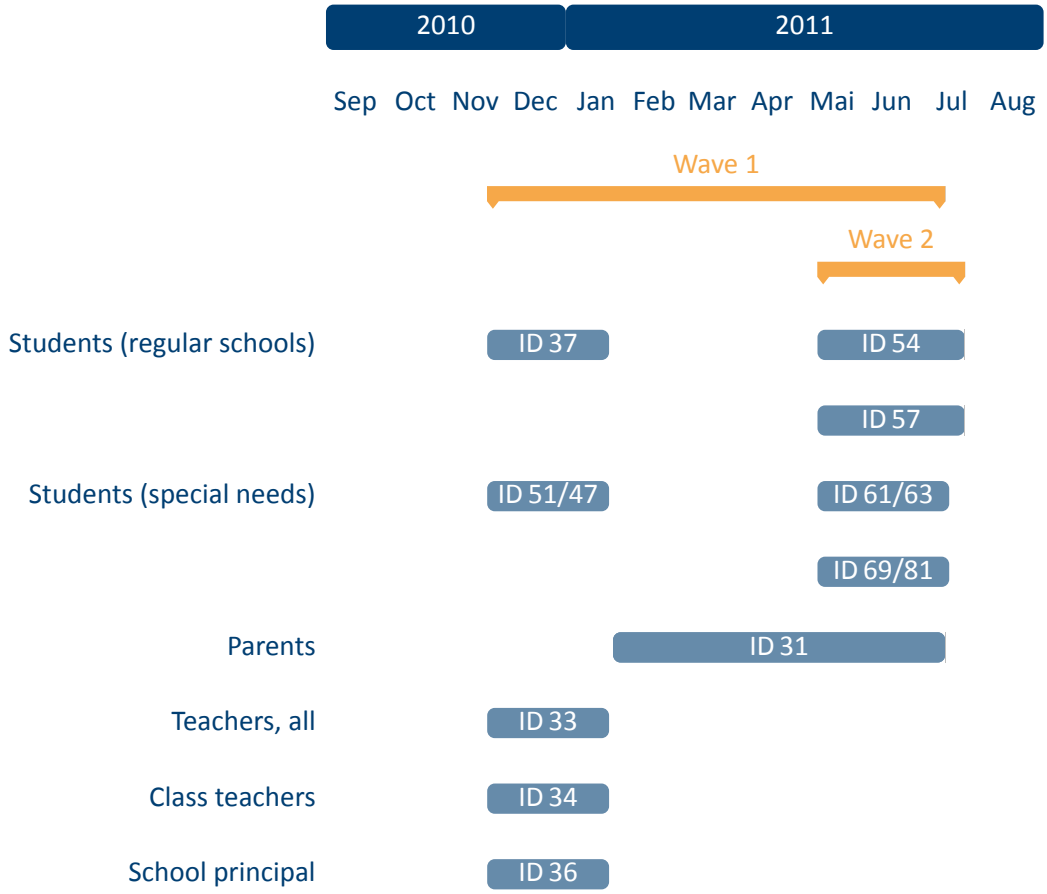


Figure 4: Order of the surveys for starting cohort 4 in survey year 2010/11 (wave 1 and 2) and IDs of the questionnaires.

vocational training in second wave. Those and the other students who left school after grade nine were surveyed individually in the context of NEPS stage 6, for the first time in fall 2011.

Teachers, principals of schools ("Schulleiter"), and parents of children were surveyed too, however only in the first wave (fall 2010). Contrary to Starting Cohort 3, only class teachers ("Klassenlehrer") were asked. Surveying of math teachers ("Mathematiklehrer") and German teachers ("Deutschlehrer") is projected in third school-context wave of starting cohort 4 (tenth grade, spring 2012). Besides a general teacher questionnaire with more personal questions there is a special questionnaire for class teachers regarding class issues. A further questionnaire was given to school principals in order to collect data about structural characteristics, equipment and the social and ethnical composition at the school. Teacher as well as principal questionnaires were self-administered paper-and-pencil questionnaires.

Similar to the Kindergarten and 5th graders cohort, a 30-minute-interview with parents has been conducted via computer assisted telephone interview (CATI) to collect comprehensive

context and background data on the target persons. This was organized in a separated study (B34) with almost parallel field time. Parents were interviewed in the first wave, not in the second. From the set of contacted parents about 82.7% did participate in the telephone survey, about 13.7% temporary dropped out (i. e., will be contacted again in subsequent parent study), and 3.6% finally dropped out (e. g., withdrew from survey participation). The field report (Aust, Hess, and Prussog-Wagner 2013, forthcoming[b]) will give more information on this survey.

Table 7 on page 23 provides an overview to respondents, modes, instruments, and contents of data collection of the first and second wave. The following sections provide an overview to sampling and survey design.

3.3 Sampling

The population of Starting Cohort 4 is defined of all students attending secondary school in fifth or ninth grade in Germany in school year 2010/2011. Access to this population is gained via the corresponding institutions. The set of secondary schools involves all officially recognized and state approved educational institutions in Germany providing schooling for fifth or ninth grade students. Students attending vocational schools or schools with a predominant foreign teaching language hindering the realization of a complete survey procedure with the available test instruments are excluded. Also, students not able to follow the normal testing procedure attending regular schools are excluded. Additionally, NEPS comprises a sample of students attending special need schools. Here the focus is put on schools with main emphasis on learning disabilities. Access to this population is gained via special need schools with federal-state-specific provisions explicitly for students with learning disabilities. Overall, 80% of children attending special need schools have a diagnosed learning disability constituting the largest group of students in these schools.

The samples for students in fifth and ninth grade are established on the basis of stratified multistage sampling designs utilizing a partly overlapping school sample. As common for surveys focusing on these age cohorts (e. g., PISA, TIMSS, TREE) the sample is drawn using an available sampling frame. On first stage schools are sampled and on second stage two classes (if available) of grade five and nine, respectively, are selected.

The large variety of federal-state-specific school systems is a challenge for sampling fifth and ninth grade students. Many different school types related to different transitions between elementary and secondary schooling institutions make up the set of schools providing access to the target population of fifth and ninth grade students. To reflect this large variety, seven forms of school have been defined as explicit strata to sample schools: Gymnasien, Hauptschulen, Realschulen, Integrierte Gesamtschulen, schools offering all tracks of secondary education except the academic track (Schulen mit mehreren Bildungsgängen), schools offering schooling to special needs children with learning disabilities, schools providing schooling only to fifth grade students, but not to ninth grade students.

For detailed discussion of the sampling design have a look at Aßmann, Steinhauer, Kiesl, et al. (2011). For details on sampling and weighting have a look at Aßmann, Steinhauer, and Zinn (2012).

3.4 Surveys and Tests

3.4.1 Competence assessments

Regular school students of Starting Cohort 4 were tested in different domains in both waves. Note, competence data is available only for regular students in this release. Due to validity concerns of NEPS pillar 1 test data of the special needs student sample is not included in this release.

Assessment in Wave 1 (fall 2010)

In first wave (fall 2010), regular students were tested in ICT literacy, scientific competence, reading speed, mathematical competence, and listening comprehension at word level (receptive vocabulary). In addition, for each of those domains procedural metacognition was tested. The test situation was organized as follows: group testing, normally taking place in the classroom, individual tests, single seats, 1 test instructor, normally 1 supervisory.

The tests are predetermined in two different sequences (random order of the test booklets for the study participants):

order test booklet 1 ICT literacy, scientific competence, reading speed, mathematical competence, listening comprehension at word level (receptive vocabulary)

order test booklet 2 scientific competence, ICT literacy, reading speed, mathematical competence, listening comprehension at word level (receptive vocabulary)

There was one 15-minute break before the test to measure reading speed. Duration of test was 112 minutes (net processing time). Learn more about the test situation of 9th grade students in regular schools by consulting the document "Main Study 2010/11 (A46), Students, 9th Grade, Regular Schools–Information on the Competence Test"[3].

Assessment in Wave 2 (spring 2011)

In second wave (spring 2011), regular students were tested in reading competence, declarative metacognition, nonverbal cognitive basic skills (perceptual speed + reasoning), listening comprehension in russian or turkish language (locator test + L1 test). In addition, domain-specific procedural metacognition was tested for reading competence and the L1-test. The test situation was organized as follows: group testing, normally taking place in the classroom, individual

3 https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/C_A46_EN.pdf

tests, single seats, 1 test instructor, normally 1 supervisory. There were two testing days. On first day all students were tested. On the second day only students from migrant families (former Soviet Union and Turkey) were participating. Tests were administered in a fixed sequence on each day:

order test booklet Day 1 reading competence + procedural metacognition, declarative metacognition, cognitive basic skills (perceptual speed + reasoning)

order test booklet Day 2 Russian or turkish locator test, listening comprehension in russian or turkish language (L1 test)

Test duration on first day was 56.5 minutes (with short breaks between different tests), on second day 34.5 minutes (with a 5 minute break after the locator test). To learn more about the test situation of 9th grade students in regular schools by consulting the document "Main Study 2010/11 (A47), Students, 9th Grade, Regular Schools–Information on the Competence Test"[4].

Documentation resources on assessment and scaling of competence data

Weinert et al. (2011) provides a comprehensive discussion on the measurement of competencies across the life span in the NEPS. For a detailed treatment of the scaling of competence data consult the technical report of Pohl and Carstensen (2012). A report on scaling results on reading competence data can be found at Pohl, Haberkorn, et al. (2012). Results to scaling of computer literacy is provided in Senkbeil and Ihme (2012). You will find details on the assessment of declarative metacognition in Lockl (2012a) and procedural metacognition in Lockl (2012b). Please cite Edele, Schotte, and Stanat (forthcoming) regarding the construction and validity of the listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish and Edele, Schotte, Hecht, et al. (2012) regarding the scaling of these instruments.

3.4.2 Student questionnaire

After testing, students received a questionnaire. This covers several topics like socio demographics, learning environments, cultural capital, educational aspirations, satisfaction, health, language use, reading engagement, or quality of instruction.

Regular students got a first questionnaire in wave 1 (fall 2010) and a panel questionnaire in wave 2 (spring 2011) if they had participated at first wave. Students in wave 2 that missed participation in wave 1 got a modified first questionnaire in wave 2. Table 6 on the next page gives an overview to the chapters of the different questionnaires. See section A.1 on page 53 for detailed information on how the questionnaires have been administered over waves.

Naturally, collecting questionnaire data from students with special educational needs ("Förderschüler") is challenging. Thus, NEPS developed different field work procedures and instruments for students from the special needs sample. For example, there were several rotations for

4 https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/C_A47.pdf (German, English forthcoming)

Table 6: Chapters of questionnaires for regular students in first and second wave

Chapter	Wave 1 (fall 2010)	Wave 2 (spring 2011)	
	First, ID 37	**First, ID 57** (if missed wave 1)	**Panel, ID 54**
1	About you	About you	About you
2	More about you	Vocational career	Vocational career
3	Family	School leaving certificate	School leaving certificate
4	Computer usage	Plans for the future	Plans for the future
5	Reading	Origin (migration)	Origin (migration)
6	Leisure time	Leisure time	Leisure time
7	School	Practical training course	Practical training course
8	Vocational career	More about you	More about you
9	Origin and language	Health	Health
10	—	Family	—
11	—	Origin (migration)	—

the student's questionnaire with different arrangements of topics (see section A.2 on page 54). Moreover, questions were read aloud by the supervisor.

3.4.3 Parent interview

To get background information about the child and the parental context, parents of the children were invited to participate in a supplementary study (NEPS study B34 for starting cohort 4). Parents of children have been interviewed via telephone (CATI mode) based on their consent that was given in advance. Preferably selected for interview were those parents who are responsible for the everyday issues of the child. In the interview data on home-learning environments, cultural and social capital, language use and proficiency, health of the target person, social origin and migration status as well as school history have been collected.

For Starting Cohort 2 (Kindergarten), 3 (fifth graders), and 4 (ninth graders), a common CATI instrument has been employed. There is a core program of questions that have been delivered to all parents of the three starting cohorts. However, there are modules which have been delivered specifically by cohort, e. g. only to parents in Starting Cohort 2 (like the child care history). Table 21 on page 60 shows which parts of the CATI questionnaire were administered to which starting cohort's parents in the first wave. In starting cohort 3 and 4, the school history of children was collected retrospectively in a spell module. For ensuring the integrity of these retrospective data a check module ("Prüfmodul") has been implemented that allowed the interviewer to identify and correct inconsistencies (like temporal overlaps or gaps) in school episodes at the time of interview.

For details on field work of the first wave parent study (study B11 for SC2, study B20 for SC3, and study B34 for SC4) please have a look in the field work report provided by infas (Aust,

Hess, and Prussog-Wagner 2013, forthcoming(b)).

3.4.4 Teacher and principal questionnaires

In the general educator questionnaire, educators answered questions about their own person. For example, the educators were asked about sociodemographics, further education, pedagogic orientations, or satisfaction.

For the class teacher there was a separate questionnaire, which contained mainly questions on structural characteristics and composition of the class.

Similarily, there was a questionnaire administered to the school principals. Besides questions to the principals own person, data on structural characteristics and composition of the school as a whole were collected using this questionnaire.

3.4.5 List of students and list of teachers

For each participating school a so called school coordinator had been defined. Those coordinators were responsible for organizing all activities within schools and classes that are necessary to realize the school survey. Mostly, school principals or well experienced teachers took on the role of the school coordinators. Among other things, school coordinators maintained for each class of the school a list of students ("Schülerliste") and a list of teachers ("Lehrerliste"). See section A.3 for an example of the "Schülerliste" and section A.4 for examples of the "Lehrerliste".

In the first list all students—whether they are participating or not participating—of a class are enumerated. Besides information on the willingness to participate (consent) basic information on students is specified (like sex and date of birth). Therefore, the "Schülerliste" provides essential methodological information (e. g., needed by the NEPS methods group for generating sampling weights) on participating as well as non-participating students, however data on students without consent is not provided in the scientific use file, of course. For participating students (i. e., parents gave consent and in starting cohort 4 additionally the student gave consent) a subset of the basic data on student from this list is provided in the datasets; regardless whether students or parents responded to questionnaires.

Using the list of teachers, the coordinators had to list all class, German, and math teachers that are teaching the classes taking part in the NEPS study. For each listed teacher the coordinator subsequently specified in which class the respective teacher teaches in which function (class, German, or math teacher, see figure 6 on page 58). Thus, a multiple assignment was possible. For such schools where German and math is organized in form of "courses" ("Kurse") instead of classes ("Klassen"), the coordinator could define one or more courses per teacher (see figure 7 on page 59).

3.4.6 Logic of assignment of students and teachers to classes

As a consequence of recording students on this list, every student is uniquely assigned to one class (at one wave). However, teachers could be assigned to different classes in different roles (e. g., class teacher in class A and math teacher in class B). Moreover, to one class there could be assigned several teachers, even in the same role. Consequently, there is a n:m relationship between classes and teachers that must be considered carefully when preparing the data. Further treatment of this issue as well as practical advices will be given in section 4.2.2 and section 7.3 below.

Table 7: Overview of respondents, instruments, modes and contents of the waves of starting cohort 4

Wave 1 (fall 2011)	Wave 2 (spring 2012)
Target Person (student) – Paper-based test	
mathematics-p[5], reading speed, vocabulary-p, sciences-p, ICT-p	domain-general cognitive function, reading-p, declarative metacognition; L1-Locator, L1-Test
Target Person (student) – PAPI	
ICT- literacy	science activities
informal LE[6] (associations)	informal LE (sport lessons, jobs, youth center, nonformal LE (internship)
sociodemographics, cultural capital, social capital, educational aspirations	sociodemographics, rational choice, social capital
sociodemographics, language use, instruction in L1 and German, segmented assimilation	sociodemographics, identity and acculturation, cultural behavior, segmented assimilation, social capital
satisfaction	satisfaction, health, expected income, family aspiration
academic self-concept, Big Five	helplessness, interest, SDQ
occupational aspirations and orientations, information strategies	occupational aspirations and orientations, information strategies, discrimination
Parent of student – CATI (31)[7]	

(...)

5 "p" means procedural metacognition
6 "H" is used for home, "LE" is used for learning environment
7 The parent interview bases on an integrated instrument serving SC 2, 3 and 4. Besides the topics listed in the table the interview contains questions about the parents' and the target persons' sociodemographic characteristics, social and migration background, language of origin and language use, siblings and household context including household income.

Table 7: (continued)

Wave 1 (fall 2011)	Wave 2 (spring 2012)
nonformal LE (private lessons), transition to job school history of target person, cultural capital, social capital,identity and acculturation, traditionalism, gender roles, segmented assimilation, health, SDQ, support in job choice	(no survey)
Teacher – class teacher questionnaire – PAPI	
structural characteristics, composition, equipment,social composition, ethnic composition,quality of instruction	(no survey)
Teacher – general teacher questionnaire – PAPI	
participation in further education, pedagogic orientations, sociodemographics, migration background, L1,occupational choice, formal qualifications, workload, cooperation, satisfaction	(no survey)
Principal – PAPI	
structural characteristics, composition, equipment, staff, quality management sociodemographics of head, social composition migration background, ethnic composition, support programs for migrant composition support programs for transition to job	(no survey)
list of students	
"Schülerliste": Basic sociodemographics to children (participants & non-participants), assignment to class	(updated list)
list of teachers	
"Lehrerliste": Assignment of teachers to classes	(no list)

4 Data Structure

4.1 Overview

Aims and scope of the NEPS surveys inevitably create complex data. The idea was to organize these data in a well structured, traceable, and user-friendly way while preserving a high level of detail in the data. Occasionally, additional variables and datasets from one or more of the original files were generated to ease preparation and analysis of the data.

Usually, all cross-sectional files are stored in wide-format (denoted with x in the filename). That is, one record represents one respondent. For instance, the file `xEducator` records data from all educators responding to the educator's questionnaire. This convention does not fully apply entirely to starting cohort 4. As students were interviewed twice, the scientific use file features two waves. The content of these two waves differs to a large extent so that their data structure is best represented in a wide format (see section 4.2.3 for a more detailed description). For episode data, usually collected retrospectively using iterative sets of questions, we provided so called spell files that are prefixed by `sp`. An example in starting cohort 4 is the file `spSchool` that contains a student's school history reported by one of her parents. Besides questionnaire and test data provided by respondents, there is also para data provided in the scientific use file.

Table 8 on the next page provides an overview to the data files contained in the scientific use file of Starting Cohort 4 (release version 1.1.0).

Note, that since school cohorts of the NEPS have a multi-level and multi-informant design there are identifiers of several units to be considered:

ID_t Identifies a target person (here: a student). `ID_t` is unique over waves and over starting cohorts.

ID_e Identifies a teacher uniquely. It can be used to match teacher data to courses directly and to match teacher data to students indirectly via their course assignments. `ID_e` is unique over waves and over starting cohorts.

ID_p Identifies the parental context of a child. `ID_p` is unique over waves and over starting cohorts. `ID_p` is needed to match data files that were generated in the parental interview. For students having parents, who gave their consent but could not be contacted in the data collection period (e. g., due to unidentifiable address data) or who did not agree to participate are not included in the parent files, there is a missing code of -55 ("not determinable").

ID_cc, ID_cg, ID_cm Identify the class, German, or mathematics course uniquely within a wave. It will not be unique over waves. Every student is assigned to any one of these three courses. Note that the dominant organization is the class rather than the course context.

This implies that a lion's share of the identifiers is identical within one wave given a student (i. e., `ID_cc` = `ID_cg` = `ID_cm` for one target). However, there are several cases where this does not apply.

ID_i Identifies a school facility uniquely. `ID_i` is unique over waves and over starting cohorts.

There are additional identifier variables for marking a child's membership to a test group (`ID_tg` in `CohortProfile`) and for marking an interviewer in the parent interview (`ID_int` in `ParentMethods`). However, these IDs are not relevant for data merging and for most empirical applications negligible.

Table 8: Summary of data files in 1st and 2nd wave of starting cohort 4 (9th Graders)

File	Content	ID[8], rows, avail.[9]
Linkage		
CohortProfile	Para data on all students of the panel sample, wave 1 and 2 (long format: 1 row = 1 student in 1 wave)	`ID_t wave` 32850 D, R, O
Survey and test data collected at school		
xTarget	Student questionnaire data, wave 1 and 2 (wide format: 1 row = 1 student)	`ID_t` 16254 D, R, O
xTargetCompetencies	Test data, wave 1 and 2 (wide format: 1 row = 1 student)	`ID_t` 15093 D, R, O
xEducator	Data from general teacher questionnaire (flat format: 1 row = 1 teacher)	`ID_e` 1651 D, R, O
xInstitution	Data from principal questionnaire (flat format: 1 row = 1 principal)	`ID_i` 493 R, O
xCourseClass	Assignment of class teachers to classes and data from class teacher questionnaire (long format: 1 row = 1 assignment)	`ID_cc ID_e` 1301 D, R, O
xCourseGerman	Assignment of teachers to German classes (long format: 1 row = 1 assignment)	`ID_cg ID_e` 1256 D, R, O

(...)

8 one row is uniquely identifiable by this variable(s)
9 availability of file: D = download, R = RemoteNEPS, O = on-site

Table 8: (continued)

File	Content	ID, rows, avail.
xCourseMath	Assignment of teachers to math classes (long format: 1 row = 1 assignment)	ID_cm ID_e 1296 D, R, O
Data from parents		
xParent	Most of data from parent interview (flat format: 1 row = 1 parent)	ID_p 9173 D, R, O
spSchool	Spell data on students' school history collected in the parent interview (spell format: 1 row = 1 spell)	ID_p spell 21768 D, R, O
spGap	Spell data on gaps in school history collected in the parent interview (spell format: 1 row = 1 spell)	ID_p spell 345 D, R, O
ParentMethods	Para data from the parent interview (includes all contacted parents, flat format: 1 row = 1 parent)	ID_p 11087 D, R, O
Regional data		
xInstitutionRegioInfas	Regional data to enhance analyses taking into account the institution (long format: 1 row = 1 institution at 1 structural level)	ID_i regio 2588 O
xTargetRegioInfas	Regional data on the parent's place of residence (long format: 1 row = 1 student at 1 structural level)	ID_t regio 44348 O
xInstitutionMicrom	Regional data to enhance analyses taking into account the institution (wide format: 1 row = 1 institution)	ID_i 647 O
xTargetMicrom	Regional data on the parent's place of residence (wide format: 1 row = 1 student)	ID_t 11087 O
Weights		
Weights	Weights are provided in a stand-alone datafile (wide format: 1 row = 1 student)	ID_t 16425 D, R, O

4.2 Data files

4.2.1 CohortProfile: para data on the cohort's panel sample

The file `CohortProfile` contains all target persons of the panel sample. In Starting Cohort 4 these are all students with agreement to participation. The file has a long format structure with variables `ID_t` and `wave` identifying a row uniquely. For each wave and student variables on sample assignment, data availability, and participation status are provided. Furthermore, data from the list of students ("Schülerliste", filled out by the school coordinators, see section 3.4.5 and section A.3) are stored in this file. In general, we strongly recommend using this file as a starting point of any analysis (see syntax examples in section 7).

The `CohortProfile` contains useful generated variables for sex and date of birth of students: Due to the study design, these very basic information was collected at several occasions. In Starting Cohort 4, information on a target's sex and date of birth is available in the student's questionnaires (wave 1 and wave 2), the parental interview, as well as the list of students filled out by the school coordinators. Thus, e.g. for date of birth there are up to 4 sources of information. Nevertheless, while not perfect there is a very high accordance between data given in the list and data provided by the students (conditional on non-missing values). However, when working with these variables one should remember the fact that they are not exactly the same since they result on different data-generation processes (data provided either by school coordinators, students, or parents). To solve this issue `CohortProfile` provides harmonized variables for sex (`tx29001`) and date of birth (`tx2900m` and `tx2900y`) which have been generated according to a hierarchical rule: (1) take as value the student's answer to the first questionnaire (mostly wave 1); (2) if last step results in a missing value, take the value from the panel questionnaire (second wave); (3) if last step result in a missing value, take the value from the parent interview; (4) if last step results in a missing, take the value noted in the list of students by the school coordinator. We treated date of birth as a coupled month-year information, i.e., we did not mix up a value for month of birth provided by the student with a value for year of birth provided by the parent or the students' list.

There are three important identifier variables that allow an unambiguous assignment to classes and courses for each student: `ID_cc` represents the identifier for the class (for merging with file `xCourseClass`), `ID_cg` for teaching in German (for merging with file `xCourseGerman`), and `ID_cm` for teaching in mathematics (for merging with file `xCourseMath`). See following section 4.2.2 for further details on class data. For mapping students to parents use the variable `ID_p` that identifies the parental context.

Table 9 on the following page presents a data example of the `CohortProfile` file. There are three students. According to the records from the list of students they are born between 1994 and 1996 (harmonized variable `tx2900y`), two students are male and one is female (harmonized variable `tx29001`, 1=male and 2=female). For student with `ID_t`=4011747 there is no assignment to parents (code -55 in `ID_p`), because his/her parents gave no consent or could not

Table 9: Exemplary data snapshot of `CohortProfile`

ID_t	wave	stratum	ID_p	ID_cc	tx80220	tx80524	tx2900y	tx29001
4010909	1	4	1012142	1000257101	1	1	1994	1
4010909	2	4	1012142	1000257101	1	-54	1994	1
4011747	1	5	-55	1000593102	1	1	1995	2
4011747	2	5	-55	1000593102	1	-54	1995	2
4011748	1	3	1010796	1000285101	1	1	1996	1
4011748	2	3	1010796	1000285101	2	-54	1996	1

be contacted in the field work. The variable stratum identifies the sample stratum (here school form) to which the student belongs. Student with `ID_t`=4011748 did not appear to school survey at wave 2 and is coded as a temporary dropout in wave 2 (`tx80220`=2). Variable `tx80524` indicated whether data on the school as provided by the principal is available. For all students school-level data is available in wave 1. Since there was no principal questionnaire in wave 2, `tx80524` is consistently set to missing code -54 (missing by design) for wave 2.

4.2.2 xCourseClass, xCourseGerman, xCourseMath: mapping of teachers and classes and data on the class context

Usually, teaching in school is organized in classes being supervised by one teacher who does not supervise other classes. However, special forms like co-teaching cause multiple assignments of classes to teachers, and vice versa. As a result, in `xCourseClass` you will find one teacher having up to two classes and one class having up to five teachers.

Furthermore, in the 9th grade teaching in German and math is usually organized in a class rather than a course context. This means that the teacher who provides information about the class will also provide information about the teaching in German. However, there is a small amount of schools in which there are courses and not classes. Therefore, in order to have a generalized and flexible data format we organized the data in a course format. Every "course" (class, German, or math) provides a unique identifier. `ID_cc` represents the identifier for the class (`xCourseClass`), `ID_cg` for teaching in German (`xCourseGerman`), and `ID_cm` for teaching in math (`xCourseMath`). Note that the identifiers are unique only within but not across waves. Assuring that a class from one year equals a class from another year is a difficult task and necessitates precise assumptions about for example the variability in the structure of the class over time.

The course files link courses to (potentially multiple) teacher(s) in a simple long format where one row corresponds to one course-teacher assignment. Note that this multiple course-teacher assignment in technical terms represents an n-to-m relationship between courses and teachers. If a teacher filled out the specific questionnaire about the respective course you will find

the given information attached to the course-teacher assignment. If not, of course you can access the assignment but the corresponding questionnaire variables contain the value -56 ("not participated"). In wave 1 of Starting Cohort 4, questionnaire information on courses is only availably in `xCourseClass`, since there was only a questionnaire specific to classes and not to German or math.

To simply this n-to-m relationship we provide a generated dummy variable `ex20100` indicating the recommended course-teacher linkage by value 1. Restricting the dataset to rows having `ex20100==1` you get a quite easy to use one-dimensional dataset with one class per row having a "default" teacher assignment (for an illustrative application see example in section 7.3. Note, that `ex20100` is generated simply by defining those teachers as the default who provide the most non-missing values within a class (if all teachers contribute equally, the first teacher according the order in the list of teachers was selected). Be aware that this is just a technical definition; in the context of your analysis it might be reasonable to rely on more substantial selection criteria–you are free to do that just by ignoring the dummy.

Table 10: Exemplary data snapshot of `xCourseClass`

ID_cc	ID_e	ID_i	ex20100	ex20101	e451010	e229400_D
1000153102	1006242	1000153	1	0	-56	-56
1000154101	1001153	1000154	1	1	8	3
1000173101	1001725	1000173	0	1	-90	-90
1000173101	1003385	1000173	1	1	-90	5
1000173101	1005787	1000173	0	0	-56	-56
1000184101	1005222	1000184	1	0	-56	-56
1000184101	1005332	1000184	0	0	-56	-56
1000184101	1005446	1000184	0	0	-56	-56
1000194103	1001854	1000194	1	1	8	5
1000194103	1006094	1000194	0	0	-56	-56

Table 10 shows an exemplary snapshot from the `xCourseClass` data file. Depicted are 10 records relating to 5 classes (or more generally "courses") identified by `ID_cc`. The educator identifier variable `ID_e` references those teachers, who were assigned as class teachers to each of the classes. Hence, each row represents a unique teacher-class relation. The school to which the class belongs is indicated by `ID_i`. As noted above, `ex20100` indicates the "default" teacher assignment as defined by the NEPS data edition. Variable `ex20101` indicates whether the teacher participated at the class teacher questionnaire. Consequently, for those cases where `ex20101` evaluates to 0 variables corresponding to items of the class teacher questionnaire contain missing values (−56). Eventually, `e451010` and `e229400_D` are two arbitrarily selected variables from the class teacher questionnaire.

The example shows that the assigned class teacher of the first class did not respond to the questionnaire at all (`ex20101`=0). Consequently, in this case all questionnaire variables are coded

to −56. This teacher is trivially assigned to be the "default" class-teacher since no other teachers have been assigned to this class. In the second case (ID_cc=1000154101) there is only one teacher assigned, too. However, this teacher provided information to the class by completing the class teacher questionnaire. This class is an "ideal" case. One teacher was assigned, who participated and provided non-missing information. The third class (ID_cc=1000173101) has three records due to the fact that three teachers have been assigned. One of those did not fill out the class teacher questionnaire while the two others did. Teacher (ID_e=1003385) is marked as the default teacher, because he provided more non-missing information to the class (missing −90 refers to item non-response). The forth class (ID_cc=1000184101) also has three teachers, but none of them responded to the questionnaire. Hence, the first teacher given the order on the list of teachers was selected as default (lowest value in ID_e).

The same logic applies for data files xCourseGerman and xCourseMath. However, remember that those files do not include questionnaire variables, since in the first wave there were no German and math teacher questionnaires.

4.2.3 xTarget: data from student questionnaires

Students are target persons in Starting Cohort 4. In first and second wave, students got different questionnaires (section 3.4.2 on page 20). Data from both questionnaires is integrated in data file xTarget. Since wave 1 and wave 2 questionnaires differ substantially with regard to asked questions the wave integration follows a "wide" concept rather than a "long" concept. As a result, xTarget follows a simple data structure with one row pertaining to one student responding to at least one of the wave 1 and 2 questionnaires. Columns contain variables collected from both surveys. Since there is one row per student the file is labeled by prefix "x" indicating a cross-sectional data structure (however, by comprising two wave data it actually contains longitudinal information).

Students are uniquely identified by ID_t. Note, that xTarget includes only those students of the panel sample who respond to at least one of the both waves. Thus, students, who did not fill out either of those school questionnaires (e. g., they did not show up at any testing day) are not included. If you want to consider the whole student sample consult the data file Cohort–Profile, which includes all students of the panel sample regardless of their questionnaire participation (section 4.2.1 on page 28).

As a result of the wide integration xTarget is organized in three blocks–basic, wave 1, and wave 2. To ease the selection of these three blocks there are two offset variables DELIMITER1 and DELIMITER2 which separate them visually from each other. The first block entails the so called basic questions. They cover mainly socio demographics and have to be administered only once. Students taking part in the survey at wave 1 were asked these questions in advance of the second block. If a student temporarily dropped out (e. g., due to illness) in the first wave, he or she has been administered these initial questions at participation in the second wave (student got a combined version of first and panel respondent questionnaire). The second block entails the questions that are dedicated to wave 1 and after the initial questions

in wave 1. The third block entails variables of the panel instrument of wave 2. Questions which were identically posed to students repeatedly in the two waves are indicated by the suffix "_w1" and "_w2" (see table 3 on page 6). A student who temporarily dropped out in one of the two waves has been assigned the value −56 (not participated) on all variables representing the questions of that particular wave. Additional to the variables indicating participation in `CohortProfile` this eases data preparation a good deal. Section A.1 on page 53 provides a comprehensive documentation of the administering of student questionnaires to participating students. Finally, the variables containing harmonized data on sex and date of birth (duplicated from `CohortProfile`, see section 4.2.1 on page 28) are also included in `xTarget`.

A slight exception to this block logic is the survey concerning students with special educational needs. Usually, their questionnaire is simply a wave-specific subset of questions posed in the regular schools. In wave 2, however, students of special schools were asked questions 61 to 64 from the regular students' questionnaire of wave 1. Of course, these questions were not posed to them in wave 1 already. What does this mean for your data preparation? If you don't especially analyze students from special schools in comparison to students from regular schools, this issue is negligible.

Table 11 shows a small snapshot from `xTarget`. There are 5 students. Dummy variable `wave_w1` (`wave_w2`) indicates whether the row contains data from wave 1 (2). Hence, first three students filled out questionnaires in both wave, whereas the last two students filled out just one questionnaire (either in wave 1 or wave 2). Variable `t514001` measures life satisfaction. Since this item was measured in both waves, a wide suffix is applied to the variable name (see naming conventions in section 2.2.1 on page 6). Hence, there is one variable containing values for wave 1 (`t514001_w1`) and one containing values from wave 2 (`t514001_w2`). Because student `ID_t`=4011107 did not participate in wave 2, there is a missing value (-56 for "not participated") in `t514001_w2`.

Table 11: Exemplary data snapshot of `xTarget`

ID_t	wave_w1	wave_w2	t514001_w1	t101000	t514001_w2	tf0021a
4010908	1	1	10	1	8	2
4010935	1	1	6	1	7	2
4011029	1	1	6	1	5	1
4011107	1	0	5	1	-56	-56
4011362	0	1	-56	1	4	2

4.2.4 xTargetCompetencies: test data of students

Six different competence tests were administered directly on the students (reading, math, reading speed, basic cognitive competencies, orthography, and procedural metacognition). The

scored results of these tests can be found in file `xTargetCompetencies` (cross-sectional format, i. e., 1 row = 1 student). Within the data file variables are arranged by domain. At the end of each domain, we provide additionally calculated metrics based on these scored results (e. g., procedural metacognition, or wle estimators). As noted in section 2.2.2 a special nomenclature guides their variable names. For an overview to which tests have been employed at which wave consult section 3.4.1 on page 19. Pohl and Carstensen (2012) provide a comprehensive report on coding and scaling of competence data.

4.2.5 xEducator: data from general teacher questionnaire

File `xEducator` provides data from the teacher's questionnaire in a simple cross-sectional format (1 row = 1 educator). A teacher is identified in the file via `ID_e`. Only those teachers responding to the general teacher questionnaire are included.

4.2.6 xInstitution: data from principal questionnaire

Answers to the principal questionnaire are stored in `xInstitution`. Analogously to `xEducator` and `xTarget` only respondents are included. Note, due to data protection issues, this file is not available in the download version of the dataset.

4.2.7 Data from the parent interview

Parents' data comprise the actual data generated by the CATI questionnaire as well as method data generated in the context of the interview (e. g., para data describing the interview). Moreover, questionnaire data consists of cross-sectional data (at time of interview) as well as of retrospective data stored in a convenient spell format. While one integrated CATI instrument has been employed for Starting Cohorts 2, 3, and 4, parent files are provided separately for each cohort. The instrument contains cohort-specific filtering so that in effect some questionnaire modules were not relevant for all cohorts. The basic identifier in all parent files is the variable `ID_p` uniquely marking a parental context (i. e., a parent).

NEPS data edition organizes data from parents' interviews in several data files. Table 22 on page 61 shows by starting cohorts which data files relating to the parent interview are available. In the following, we describe those parent files from wave 1 of starting cohort 4.

`xParent`

Most of the data from the parent interview is stored in `xParent`. It has a simple cross-sectional data structure with one row corresponding to one interviewed parent identified by `ID_p`. Only data from realized interviews is included.

spSchool

In Starting Cohort 4, the parents are surveyed retrospectively on their child's schooling history. Since this data follows an episode structure it is stored in a separate file named spSchool. Therefore, spSchool has a long data structure with one row marking a specific school episode reported by a specific parent (see data example in table 12). Thus, a row is uniquely identified by ID_p and the counter variable spell. Each spell has time variables: a start date and an end date each of which coded by month and year (p72301m/y for start and p72302m/y for end). If the reported end month equals the month of interview a censoring question asks whether the school episode is still lasting or ended within the month of interview (p723110). Reported start and end dates have been checked for inconsistencies and wrong temporal ordering at the time of interview by a so called check module (see section 3.4.3). In the check module the interviewer was able to record corrected dates. These corrected spell times are available in the generated variables *_g1. For example, p72301y_g1 contains the year of the episodes start date corrected by the check module. While we generally recommend using the corrected times, the decision whether you want to work with the originally reported or the corrected times is up to you as a researcher.

Table 12: Exemplary data snapshot of spSchool

ID_p	spell	p723080	p72301m_g1	p72301y_g1	p72302m_g1	p72302y_g1	p723110_g1
1002312	1	1	8	2002	8	2006	2
1002312	2	4	8	2006	2	2010	2
1002312	3	4	4	2010	8	2010	2
1002312	4	4	8	2010	2	2011	1
1002313	1	1	8	2003	7	2006	2
1002313	2	5	8	2006	2	2007	2
1002313	3	10	2	2007	1	2011	1
1002332	1	1	8	2002	7	2006	2
1002332	2	5	7	2006	1	2011	1
1002364	1	1	8	2002	6	2006	2
1002364	2	8	8	2006	1	2011	1

Table 12 lists data on school history records of four students as reported by their parents (ID_p) in the interview. Records represents an school episodes (identified by spell counter spell) which are nested in reporting parents. Variable p723080 indicates the form of school and the *_g1 variables contain start and end dates of schooling episodes and an additional indicator for episode censoring at time of interview (p723110_g1, if 1 then school episode is right censored at time of interview).

spGap

Contains gap episodes that have been reported in the check module dealing with the consistency of school history. For example, gaps were filled up if the student was ill or was away for

holidays. Since both contain spells in a long format, the data structure of file spGap is analog to file spSchool. Table 13 provides a snapshot of spGap. The first parent reported two gaps of different type (ps29101) in her child's school history. The two other parents in the example reported just one gap episode.

Table 13: Exemplary data snapshot of spGap

ID_p	spell	ps29101	ps2911m_g1	ps2911y_g1	ps2912m_g1	ps2912y_g1
1007613	1	12	8	2001	7	2003
1007613	2	11	8	2003	9	2003
1007615	1	11	7	2008	9	2008
1007620	1	12	8	2002	8	2003

ParentMethods

This dataset offers rich methodological information to the data collection of the parent interview (CATI mode). In particular, interviewer data (e. g., age, gender and education of the interviewer) as well as interview data (e. g., date and duration of interview, change of respondent, number of contact tries) is available.

Importantly, ParentMethods contains all contacted parents whether an interview was realized or not. Thus, ParentMethods includes more cases than xParent. A significant amount of the surplus parents in ParentMethods are just temporary dropouts (they are approached again in the second wave), but there are also some parents who finally dropped out because they withdrew their participation at the time of phone contact. You can use variable px80220 for analyzing the participation status of parents. Table 14 on the following page provides an exemplary data snapshot from the ParentMethods file. It shows some para data—state of participation(px80220), the ID of the interviewer (ID_int), the length of the interview in minutes (px80209), and sex of the interviewer (px80301)—on four parents. For the first two parents interviews have been realized (px80220=1). For the third case an interview could not be realized for some reason. Hence, this is coded as a *temporary dropout* (px80220=2). In the fourth case no interview was realized and willingness to participate was withdrawn by the parent. As a result this cases is coded as a *final dropout* (px80220=3). Of course, date and length of interview is missing for the dropout parents.

For convenience, the variable ID_t is included for mapping parents to targets, i. e. students. Note, that the mapping of children to parents—ID_t to ID_p—is already established in CohortProfile. Due to technical reasons, each ID_t is assigned an own ID_p. Consequently, siblings in the sample of students have different values in ID_p since they have different values for ID_t. From parents with more than one child taking part at NEPS child-specific questions have been surveyed in loop while context data was copied (for details see field work report Aust, Hess, and Prussog-Wagner 2013, forthcoming(a)). Hence, siblings have different rows

Table 14: Exemplary data snapshot of ParentMethods

ID_p	ID_t	wave	intm	inty	px80220	ID_int	px80209	px80301
1002312	4020001	1	2	2011	1	1157	42.46667	1
1002318	4018735	1	2	2011	1	1169	44.75	1
1018169	4023543	1	.	.	2	1028	.	2
1018172	4023758	1	.	.	3	1131	.	2

with a different ID_p in the data. Variable px80213 indicates whether the interview was realized in a multiple children mode.

4.2.8 Regional data

RegioInfas

Fine-grained regional data is provided in the data files xTargetRegioInfas and xInstitutionRegioInfas. These files have been generated from the *infas geodaten*[10] database. Both comprises geographical information on four regional levels (coded in variable regio): municipality, postal code, quarters (living areas), and street sections. These data were linked to targets (via the address of parents) and institutions by geocoding of the sample addresses. Regional data from xTargetRegioInfas can be easily linked to targets by selecting first the regional level (variable regio) and second using ID_t for file merging. Analogously, use ID_i for merging regional data from xInstitutionRegioInfas to institutions. Note that these data are highly sensitive and thus can only be accessed via on-site data usage. A comprehensive documentation of this dataset is provided by Koberg (2012b).

Microm

Similar data is available from *microm Consumer Marketing*[11], which is analogously provided for the same addresses of targets (xTargetMicrom) and institutions (xInstitutionMicrom) (see above). Unlike infas geodata, microm data only holds information from one regional level, which is always the most fine grained available. You may spot this level by using variable ebene. Benefit of this is that you only need one variable (ID_t or ID_i) for unique identification and merging. Microm data also can only be accessed on-site.

10 This is a data product of infas geodaten GmbH, see http://www.infas-geodaten.de.
11 http://www.microm-online.de/

4.3 Syntax for cleaning filtered question data (PAPI)

Of course, filtering in self-administered questionnaires does not perfectly work. As a consequence, respondents sometimes give answers to questions which they should actually skip according to the responses they gave before (check questionnaires for identifying the respective filters). However, if they do so, one can hardly determine *ex post* which given information is valid and which invalid (the filter question was wrongly answered or the questions to skip were wrongly answered). Therefore, NEPS data edition does not manipulate on the basis of uncheckable assumptions, however provides cleaning syntax to account for that problem. In general, following a linear principle, i. e. assuming the first response is the valid one, the cleaning syntax recodes a −99 missing code (filter missing) for such variables that actually should have been skipped by the respondent. We only provided cleaning syntax for filters in the context of paper questionnaire data.

The cleaning syntax can be downloaded from the SC_4 data section on our web portal (see section 10). If you are using RemoteNEPS or the on-site version of the SUF, you additionally will find the syntax files in the directory of your SUF data. Names of the cleaning files are almost identical to the data files, differing only in their filename extensions. For Stata their names are `*.do` and for SPSS their names are `*.sps`.

To get a quick understanding of what happens when you apply the cleaning syntax, you should read the `readme_filtering.txt` file in the same directory. It describes the cleaning procedure in detail. Also see example 5 in section 7.4 for an application of the cleaning syntax.

To avoid errors, the cleaning syntax should be started before you do any recoding of the data.

5 Coding

Occupational strings (like respondents' favored jobs, the desired vocational training, and the idealistic and realistic occupational aspirations as well as the parents' occupations and many more) were coded and several classifications and schemes were derived. Table 15 on the next page presents an overview of these coded variables. Furthermore, the parental educational information was coded by using the CASMIN and the ISCED-97 classification and a metric variable offering the standardized years of occupation. See table 16 to table 19 on pages 39–41 for a documentation on how EGP (cf. Erikson, Goldthorpe, and Portocarero 1979), the BLK (classification of occupations according to Blossfeld 1985; Schimpl-Neimanns 2003) the ISCED-97 (UNESCO 2006) and CASMIN (Lüttinger and König 1998) classes are coded.

Table 15: Overview of coded variables

Classification	Included in	Description
KldB88	`xEducator`; `xParent`	German Classification of Occupations 1988 (4-digit)
KldB2010	`xEducator`; `xParent`	German Classification of Occupations 2010 (5-digit)
ISCO-88	`xEducator`; `xParent`	International Standard Classification of Occupations 1988 (4-digit)
ISCO-08	`xEducator`; `xParent`	International Standard Classification of Occupations 2008(4-digit)
BLK	`xEducator`; `xParent`	Occupational classification by Blossfeld based on KldB92 (cf. Blossfeld 1985; Schimpl-Neimanns 2003)
ISEI-88	`xEducator`; `xParent`	Metric scale to measure the socio-economic status of occupations based on ISCO-88 (cf. Ganzeboom et al. 1992; Ganzeboom 2010)
ISEI-08	`xEducator`; `xParent`	Metric scale to measure the socio-economic status of occupations based on ISCO-08 (cf. Ganzeboom et al. 1992; Ganzeboom 2010)
SIOPS-88	`xEducator`; `xParent`	Metric scale to measure prestige of occupations based on ISCO-88 (cf. Treiman 1977)
SIOPS-08	`xEducator`; `xParent`	Metric scale to measure prestige of occupations based on ISCO-08
MPS	`xEducator`; `xParent`	Magnitude prestige score of occupations (cf. Wegener 1985)
EGP	`xParent`	Class scheme which assigns occupations to classes
CAMSIS	`xParent`	Classification to measure social interaction and stratification
CASMIN	`xParent`	Classification representing differentiated educational attainment and vocational training degrees
ISCED-97	`xParent`	Classification representing differentiated educational attainment and vocational training degrees
Years of education	`xParent`	Years of education based on the CASMIN classification

Table 16: Coding of EGP

Key		English	German
1	[I]	Higher Controllers	Obere Dienstklasse
2	[II]	Lower Controllers	Untere Dienstklasse mit hohen Qualifikationen
3	[IIIa]	Routine Non-manual	Angestellte der ausführenden nichtmanuellen Klasse mit beschränkten Entscheidungsbefugnissen
4	[IIIb]	Lower Sales-Service	Angestellte der ausführenden nichtmanuellen Klasse mit gering qualifizierten Routinetätigkeiten
5	[IVa]	Selfemployed with employees	Selbständige mit unterstellten Mitarbeitern
6	[IVb]	Selfemployed no employees	Selbständige ohne unterstellte Mitarbeiter
7	[IVc]	Selfemployed Farmer	Selbständige in der Landwirtschaft
8	[V]	Manual Supervisors	Arbeiter, Techniker, Facharbeiter
9	[VI]	Skilled Worker	Qualifizierte Arbeiter
10	[VIIa]	Unskilled Worker	Unqualifizierte Arbeiter
11	[VIIb]	Farm Labor	Landwirte

Table 17: Coding of BLK

Key		English	German
1	[AGR]	Agricultural occupations	Agrarberufe
2	[EMB]	Common manual occupations	Einfache manuelle Berufe
3	[QMB]	Skilled manual occupations	Qualifizierte manuelle Berufe
4	[TEC]	Technician	Techniker
5	[ING]	Engineer	Ingenieure
6	[EDI]	Common services	Einfache Dienste
7	[QDI]	Skilled services	Qualifizierte Dienste
8	[SEMI]	Semiprofessions	Semiprofessionen
9	[PROF]	Professions	Professionen
10	[EVB]	Common commercial and adminstrative occupations	Einfache kaufmännische und Verwaltungsberufe
11	[QVB]	Skilled commercial and administrative occupations	Qualifizierte kaufmännische und Verwaltungsberufe
12	[MAN]	Manager	Manager

Table 18: Coding of ISCED-97

Key		English	German
0	0A/1A	Inadequatly completed general education	kein Abschluss
1	2B	Lower general education	Haupt-, Volksschulabschluss, Berufsvorbereitende Maßnahme
2	2A	Intermediate general education	Mittlere Reife, Realschulabschluss
3	3A	Full maturity certificates (e.g., the Abitur, A-levels)	Fachhochschulreife, Hochschulreife
4	3B	Basic vocational training, Vocational full time school, Health sector school (less than two years), civil servant of the lower grade, vocational basic skills	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse
5	3C	Civil servants of the medium grade	Beamter mittlerer Dienst
6	4A	Full maturity certificates (e.g., the Abitur, A-levels) (second cycle)	Fachhochschulreife, Hochschulreife (second cycle)
7	4B	Basic vocational training, Vocational full time school, Health sector school (less than two years), civil servant of the lower grade, vocational basic skills (second cycle)	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse (second cycle)
8	5B	Diploma (vocational and other specialised academies, College of public administration), Qualification of a two or three year Health-Sector School, Master's/technician's qualification	Fach- und Berufsakademische Abschluss, Verwaltungsfachhochschule, Fachschule des Gesundheitswesens (mindestens zwei Jahre), Meister/Techniker, anderer Fachschulabschluss, Beamter gehobener Dienst
9	5A	Bachelor, Master, Diploma, state examination, civil servants of the highest grade	Bachelor, Master, Diplom, Magister, Staatsexamen, Beamter höherer Dienst
10	6	Doctoral degree and postdoctoral lecture qualification	Promotion

Table 19: Coding of CASMIN

Key		English	German
0	1a	Inadequatly completed general education	Kein Abschluss
1	1b	General elementary education	Hauptschulabschluss ohne berufliche Ausbildung
2	1c	Basic vocational training above and beyond compulsory schooling	Hauptschulabschluss mit beruflicher Ausbildung
3	2b	Intermediate general education	Mittlere Reife ohne berufliche Ausbildung
4	2a	Intermediate vocational qualification, or secondary programmes in which general intermediate schooling is combined by vocational training	Mittlere Reife mit beruflicher Ausbildung
5	2c_gen	General maturity: Full maturity certificates (e.g., the Abitur, A-levels)	Hochschulreife ohne berufliche Ausbildung
6	2c_voc	Vocational maturity: Full maturity certificates including vocationally specific schooling or training	Hochschulreife mit beruflicher Ausbildung
7	3a	Lower tertiary education: Lower level tertiary degrees, generally of shorter duration and with a vocational orientation	Fachhochschulabschluss
8	3b	Higher tertiary education: The completion of a traditional, academically orientated university education	Universitätsabschluss

6 Weights

Weighting variables are included in the `Weights` dataset. Given the quite complex structure of the sample of school cohorts (1st wave), no final recommendations are at hand concerning the use of design and adjusted weights. Although, there are no general results available how the use of design or adjusted weights render any possible analysis (see Rohwer 2011 for a general discussion), the use of weights may possibly help to highlight important features of the analysis under consideration not at least serving as a robustness check for the performed analysis. Note that also standardized weights with mean one are provided, which are often used in statistical analysis. Also, replication weights for balanced repeated replication (BRR) are available. Information on weight construction and how to use them can be found in the technical report on weighting (Supplement C) and the examples section (see section 7.5).

7 Examples

This section gives some examples of how to work with the different data sets. We provide you with the code to run the examples in Stata.[12] In future releases of this manual we will extend the examples by code in R and SPSS.

12 In our Stata examples we make use of the user-written command `fre` which is in our view a powerful alternative to `tab` for displaying one-way frequency tables. It can be installed very easily by typing `ssc install fre` into the Stata console. Acknowledgment goes to Jann (2007), who developed this nice Stata module.

7.1 Example 1—First steps using CohortProfile

In the first example, we will have a brief look on the data using CohortProfile as a starting point. We will check frequencies of participation, sex, and year of birth of students. Furthermore, we are checking data availability for a potential analysis combining test data and data from parents.

Syntax example 1: First steps using CohortProfile in Stata

```
/*********************************************
* Example 1: Using CohortProfile

Procedure:
1. drop cases of sample of special needs students
2. check data availability of the sample of regular students

(note: replace ${version} to your file version)
*********************************************/

* load CohortProfile
use "SC4_CohortProfile_${version}.dta", clear

* switch dataset's language
label language en

* recode NEPS missings to Stata extended missings in all variables
nepsmiss _all

* drop sample of students with special educational needs in the area of learning
drop if stratum==7

* count cases in sample of regular students (for each wave separately)
bysort wave: count

* Check status of survey participation by wave
bysort wave: tabulate tx80220

* Exploit basic information on students in CohortProfile
* gender (harmonized variable)
bysort wave: tabulate tx29001
* year of birth
bysort wave: tabulate tx8050y

* check data availability: test data and parent data in wave 1
tabulate tx80522 tx80523 if wave==1, cell missing
```

7.2 Example 2—Merge datasets to CohortProfile

Taking CohortProfile as a starting point of data preparation other cross-sectional files can easily be merged. In the example shown below we merge (a) data from the target and (b) data provided by the student's parent.

Syntax example 2: Merge datasets to CohortProfile using Stata

```
/*********************************************
* Example 2: Merge datasets to CohortProfile

Procedure
1. start by loading CohortProfile
2. directly match information from
   a) xTarget and
   b) xParent dataset
3. analyse combined data

(note: replace ${version} to your file version)
*********************************************/

* load CohortProfile as master dataset
use "SC4_CohortProfile_${version}.dta", clear

* only keep observations from wave 1
keep if wave==1

* merge variable t700031_w1 (gender reported by student in first wave) from file
  xTarget
* and keep only matched cases (that is, cases who are in both files xTarget and
  CohortProfile)
merge 1:1 ID_t using "SC4_xTarget_${version}.dta", nogenerate ///
  assert(match master using) keep(match) keepusing(t700031_w1)

* merge variable p700010 (student's gender reported by parent) from file xParent
* consider missing parent IDs for some children
nepsmiss  ID_p //recode NEPS missings to Stata extended missings
* note: since there are missings in ID_p there is a m:1
* relationship for merge
merge m:1 ID_p using "SC4_xParent_${version}.dta", nogenerate ///
  assert(master match) keep(master match) keepusing(p700010)

* recode missings
nepsmiss tx80501 t700031_w1 p700010

* check agreement on student's gender between data from list of students
* (tx80501 in CohortProfile) and data from the student's questionnaire
* (t700031 in xTarget, available only for responding students, of course)
tabulate tx80501 t700031_w1, cell // >99% agreement

* agreement between students' and parents' responses on sex
tabulate t700031_w1 p700010, cell // >99% agreement
```

7.3 Example 3—Merging class data to students

As the data provide detailed and comprehensive information on the structure of the institutional context the respondents are embedded in, you might want to incorporate this information in your analysis. Say you want to study the variation of some student-level outcome variable conditional on variables on the level of classes.

In a first step you start with the `CohortProfile` and merge the student's information entailed in `xTarget`. As `CohortProfile` provides the student's course affiliation, you can easily merge the class characteristics—after you have generated them according to your specific research question. To do so, you have to choose at least one type of course. Let's start with the dataset `xCourseClass` (containing data on classes provided by the class teacher(s)). As any one of the three course datasets it reflects the n:m relationship between teachers and courses, too: One course can consist of multiple teachers and one teacher can be related to more than one courses. In this example we generate a class level variable size of class and gender of class teacher by evaluating responses from the "default" teacher (see section 4.2.2 on page 29). Subsequently, these generated variables are merged to students in `CohortProfile`.

Syntax example 3: Merging class data to students

```
/*************************************************************
* Example 3: Indirectly match teachers to students
                        by preparing course-level data
Procedure
1. prepare course-level data
2. load CohortProfile
3. merge prepared class-level data
4. analyse data

(note: replace ${version} to your file version)

*************************************************************/

* replace path with your local file path to the data files

* 1) Prepare course-level data:
* "class size and gender of class teacher"
* load xCourseClass file
use SC4_xCourseClass_${version}, clear

* note: ID_cc & ID_e uniquely identifies a row!
isid ID_cc ID_e

* keep default teacher assignment only
keep if ex20100==1

* now ID_cc is unique (while this is not neccessarily true for ID_e )
isid ID_cc
```

```
* merge gender of assigned class teacher from xEducator
merge n:1 ID_e using SC4_xEducator_${version}, ///
          nogen keep(master matched) keepusing(e762110)
rename e762110 gendteach

* class size (take collapsed version)
rename e227400_g1D clsize

* keep only relevant variables
keep ID_cc ID_e gendteach clsize
* recode missings
nepsmiss gendteach clsize

* temporarily save prepared class file
tempfile classcontext
save ‘classcontext’, replace

* 2) Load CohortProfile file
use ID_t ID_cc wave stratum using SC4_CohortProfile_${version}, clear

* restrict dataset to first wave (fall/winter 2010)
keep if wave==1

* 3) Merge class-level data
*  (note: due to nonresponse on the teacher level
*  information is not available for all classes)
merge n:1 ID_cc using ‘classcontext’, nogen keep(master matched)

* 4) Analyse data (e.g., inspect distribution of class size
*                                  and gender of teacher over children)
tab clsize, miss
tab gendteach, miss
```

7.4 Example 4—Using cleaning syntax for accounting filtering in paper and pencil questionnaires

In the example we do filter cleaning for the data from teachers. See section 4.3 for details on using the filter question. Basically, data cleaning syntax files have the same name as the corresponding data file. Make sure, that the version number of the cleaning syntax matches the data version.

Syntax example 4: Merging class data to students

```
/*********************************************
* Example 4: Using filtering syntax

Note: replace ${version} to your file version
*********************************************/

* load xEducator
use SC4_xEducator_${version}, clear

* e.g. variable e537042
* Before: missing codes -95 (implausible) and
*         -90 (unkown missing)
fre e537042

* apply filter syntax
* (assuming do-file stored in the same directory like the data file)
* (note: make sure that the filter syntax version number matches
*        your data version!)
quietly do "SC4_xEducator_1-0-1.do"

* e.g. variable e537042
* After: missing codes -95 (implausible),
*        -90 (unkown missing), and
*        -99 (filtered)
fre e537042

* recode missings
nepsmiss *
```

7.5 Example 5—Using weights

This example intends to illustrate how weights provided in `Weights` can be used. Weights are available for different groups of respondents. In the example, we like to weight students who participated in both waves and have information from their parents available. After merging variables from `xTarget` and `xParent` to a modified `CohortProfile`, we calculate means of this variable over gender groups. Two ways of weighting are demonstrated: One manually setting the weighting mechanism, and the other one using Stata's powerful –svy– command. The latter variant incorporates clustering of standard errors for primary sampling units where appropriate.

Syntax example 5: Using weights

```
/*********************************************
* Example 5: Using weights

Note: replace ${version} to your file version
*********************************************/

* load CohortProfile
use "SC4_CohortProfile_${version}", clear

* first we have to trim the CohortProfile so it only includes
* those variables and cases we want to observe
keep ID_t ID_p wave tx80220 tx80523 stratum

* reshape "survey subject available" and "survey parents available"
reshape wide tx80220 tx80523, i(ID_t) j(wave)

* only keep relevant cases, which is:
* a) students who have participated in both waves and
* b) also have information from their parents available
keep if (tx802201==1 & tx802202==1) & tx805231==1

* now we are ready to merge the corresponding weight from Weights
merge 1:1 ID_t using "SC4_Weights_${version}", assert(match) keep(master match)
  keepusing(ID_i w_tp1_std) nogenerate

* fetch some interesting variables from xTarget ...
merge 1:1 ID_t using "SC4_xTarget_${version}", assert(master match) keep(master match)
  keepusing(t700031_w1 t66000a_g1) nogenerate

* ... and also some information from xParent
merge 1:1 ID_p using "SC4_xParent_${version}", assert(using match) keep(match)
  keepusing(p261100) nogenerate nolabel

* encode missings
nepsmiss t66000a_g1 t700031 p261100 ID_p

* calculate means without weighting
mean t66000a_g1, over(t700031_w1 p261100)
```

```
* calculate means using probability weights
mean t66000a_g1 [pweight=w_tp1_std], over(t700031_w1 p261100)

* calculate means using sampling weights (svy-command)
svyset ID_i [pweight=w_tp1_std], strata(stratum)
svy: mean t66000a_g1, over(t700031_w1 p261100)
```

8 Rules and Recommendations

8.1 Rules

Always remember the rules and stipulations that you have agreed when signing the NEPS data contract! In particular:

- remember, that, in the context of the NEPS data from Starting Cohort 2, 3, and 4, you are not allowed to publish any analyses that aim for or allow a direct comparison of the German Bundesländer. Any forms of "rankings" of German Länder using the NEPS data is strongly prohibited. Also singling out particular Bundesländer for analyses is prohibited!
- keep secret the NEPS data provided!
- keep secret transmitted access codes (e. g. individual identification and password)!
- refrain from any action aimed at and suitable for re-identifying persons, households or institutions (e. g. education or support facilities)!
- refrain from mixing the data, and neither partially, with other data permitting the reidentification of persons!
- immediately inform NEPS of any accidental reidentication and keep secret individual data gained therefrom!

Remember, violations of stipulations and rules of the data usage contract will lead to severe penalties that are defined in the contract!

If you are not sure regarding any rule, please contact the NEPS userservice (see section 10). Also, if you encounter any security leaks regarding data protection and data security, or any data quality deficiencies please inform the NEPS userservice (see section 10).

8.2 Recommendations

We strongly recommend you to examine the data critically when you work with this release. While the NEPS invested a lot to ensure the integrity of the provided data, the latter can not be perfectly guaranteed. Furthermore, you should always consult the questionnaire/s to obtain a precise understanding of how the data have been collected. Finally, we would like to give some basic recommendations for working with the data:

- always be critical when working with empirical data!
- if you are working with Stata install and update the `nepstools` (see section 9)!
- recode missing values adequately to your statistical software!
- use file `CohortProfile` as a starting point!
- For analyses separate students with special educational needs ("Förderschüler") and regular students ("Regelschüler")!
- check documentation material and survey instruments that can be downloaded on the NEPS website (see section 10)!
- if you encounter problems or even errors in the data please contact the userservice of the NEPS (see section 10)!

9 Tools for Stata Users

Our Stata files offer variable and value labels both in German and English language. You can easily switch between these languages using the label language command: `label language en` activates English labels, and `label language de` switches back to German labels, respectively.

Furthermore, we have developed two Stata programs (`ado` files) to ease work with our data. You can obtain these files from our repository using the following command:

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

We try to fix bugs in the `nepstools` as soon as possible. Thus, you should make sure that you have the most recent version by executing following command:

```
adoupdate nepstools, update
```

nepsmiss: Recoding missing values

This program automatically recodes and labels all missing values into extended missing values (`.a`, `.b`, etc.). In this example, we run nepsmiss on the variable `t700031_w1`, decoding all negative values (like -95, -92, -90) into Stata's extended missings (like `.e`, `.i`, `.j`).

```
nepsmiss t700031_w1

t700031_w1 -- Gender child
-------------------------------------------------------------------------------
                              |      Freq.    Percent      Valid       Cum.
------------------------------+------------------------------------------------
Valid   1  ... male?          |       7853      48.31      50.52      50.52
        2  ... female?        |       7692      47.32      49.48     100.00
        Total                 |      15545      95.64     100.00
Missing .e Implausible value  |          3       0.02
        .i Unspecific missing |         29       0.18
        .j Not participated   |        677       4.17
        Total                 |        709       4.36
Total                         |      16254     100.00
-------------------------------------------------------------------------------
```

We generally recommend running `nepsmiss` on all variables (`nepsmiss _all`) before any further data preparation.

infoquery: Display survey questions

This program displays the survey question that corresponds to a variable in a dataset. Note that infoquery will produce no output for some derived variables.

```
infoquery t700031_w1
--------------------------------------------------
query result for variable t700031_w1:

t700031_w1[questiontext_en]:
Are you...
--------------------------------------------------
```

10 Further Information

Please visit our web portal for further information and comprehensive documentation resources such as PAPI and CATI questionnaires, how-to guides, technical reports, and the codebook.

→ www.neps-data.de > Data Center > Research Data > Starting Cohort 9th grade

For further support, please contact the NEPS data center:

E-mail: userservice.neps@uni-bamberg.de
Web: → www.neps-data.de > Data Center > Contact the Data Center
Phone: +49 951 8633511

Participation in the NEPS user trainings

Furthermore, the NEPS data center offers training courses on a regular basis. These courses introduce the research design of the NEPS, the structure of datasets, terms and conditions of data usage, issues of privacy and data protection, and so on. A central module of the courses consists of hands-on work with the NEPS data supervised by our staff. As skill levels, research interests, and methods vary greatly across users and disciplines, we will offer a comprehensive portfolio of seminars ranging from introductory topics on a rather general level to advanced methodological courses.

A Instruments

A.1 Administering of student questionnaires in wave 1 and 2

In wave 1 all students received a first respondent questionnaire. In wave 2, those students who already had participated at the wave 1 questionnaires got a panel respondent questionnaire. Those, who had not participated received a modified first respondent questionnaire, which contained mixed questions from wave 1 (first) and wave 2 (panel). However, in the field phase this worked not perfectly, as you can see in table 20. Few regular students ($N = 40$) who already received a first respondent questionnaire in wave 1 erroneously got the first respondent questionnaire in wave 2. Other regular students ($N = 66$) who did not participated at wave 1 questionnaire actually should had received the first respondent questionnaire in wave 2 but the panel questionnaire was delivered. Some students of the special needs sample who had not participated at wave 1 questionnaires erroneously got a panel questionnaire in wave 2 ($N = 10 + 13 = 23$).

Table 20: Administering of questionnaires to students over wave 1 and wave 2; instrument IDs in parentheses

					Wave 1				
					not. part.	regular	special needs		Total
							split 1 [47]	split 2 [51]	
						[37]			
		not participated			0	956	65	100	1121
	regular	panel		[54]	66	13494	0	0	13560
		first		[57]	543	40	0	0	583
Wave 2	special needs	panel	1a	[61]	10	0	98	410	518
			2a	[63]	13	0	366	48	427
		first	1b	[69]	25	0	0	0	25
			2b	[81]	20	0	0	0	20
		Total			677	14490	529	558	16254

A.2 Characteristics of the survey of students with special educational needs

A.2.1 General remarks on the survey of students with special educational needs

The survey of students with special educational needs (SEN) was partly different from the regular students' survey in terms of design, content, and mode. The data from SEN students has a somehow experimental nature since a lack of evidence regarding the question to which extend data from students with special needs can be adequately compared with regular students' data. In general, the questionnaire constitutes an excerpt of the tool used at regular schools. SEN students were tested and surveyed on two days. To avoid validity problems due to possible deficits in reading the questionnaire was read aloud by the testing supervisor using a predefined scriptum. Moreover, to account for possible sequence effects in responding as a result of fatigue different rotations of the questionnaires have been employed.

However, when working with the data from students with special educational needs in the field of learning it must be specifically considered that particular questions may not be entirely understood and/or the alertness and willingness to fill out a long questionnaire might not be given by all participants alike.

A.2.2 Starting Cohort 4

In wave one of the ninth graders cohort, the questionnaires for students of the SEN sample were shorter than the questionnaire of the regular students' sample. Moreover, there were two versions of the questionnaire with different order of the questions (ID 51 and ID 47). Details of question sequence and comparison of questionnaires for regular students (ID 37) and students with special educational needs is given in German National Educational Panel Study (NEPS) (2013).

The situation in Wave 2 (spring 2011) is nearly the same besides the fact that there are students, who did not answer the questionnaire in Wave 1. Those got a questionnaire with some additional questions on socio-demographics. Again there were two versions with different sequence of questions. Alltogether there were 4 different questionnaires for students of the SEN sample in Wave 2. A correspondence between the six questionnaires used in Wave 2 can also be found in German National Educational Panel Study (NEPS) (2013).

A.3 "Schülerliste"—List of students

The list of students was devided into a left, a middle and a right section. While the left section is kept at the schools for assigning students' names to IDs, the middle section and the right

section is needed for methodological purposes. The middle section contains anonymized data (without ID assignment) on participation. Since on the middle section all sampled students—whether participating or not—are listed, the methods group of the NEPS uses this section for doing selectivity analyses as well as for calculating non-response adjusted weights for students with characteristics available in the list. On the right section of the list, information only for students participating in the NEPS is available. Part of this information is available in the dataset `CohortProfile`. Note, that also the student's assignment to classes is achieved by the list, since there is a list for each class of the sample. An exemplary list is displayed in figure 5 on the following page.

Figure 5: List of students

Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 9

Schulname: Musterschule

Schul-ID: 255201
Klasse: 9c
Testgruppen-ID: 25520131

Haupterhebung K5/K9
Angaben zur Grundgesamtheit

Information im Klartext: Bundesland, Schulform, Jahrgang

Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 9

Schulname: Musterschule

Schul-ID: 255201
Klasse: 9c
Testgruppen-ID: 25520131

[1] Lfd. Nr.	[2] Name des Schülers (Nachname, Vorname)	[3] Schüler-ID	[4] Staatsange-hörigkeit 1= deutsch 2= nicht-deutsch	[5] Deutsch als vorwiegend zu Hause gesprochene Sprache (ja/nein)	[6] Noten aus dem letzten Zeugnis in Deutsch	[6] Mathe-matik	[7] Sonderpäd. Förderbedarf (ja/nein)	[8] Diagnostizierte Lese-Rechtschreib-schwäche (ja/nein)	[9] Geschlecht (m/w)	[10] Geburtsmonat und -jahr (MM.JJ)	[11] Übertrag aus Spalte 12+13 Teilnahme-bereitschaft (ja/nein)	[12] Elterng-enehmigung (ja/nein)	[13] Schülerein-willigung (ja/nein)	[14] Übertrag aus Spalte 9 Geschlecht (m/w)	[15] Übertrag aus Spalte 10 Geburtsmonat und -jahr (MM.JJ)	[16] Übertrag aus Spalte 6 Noten aus dem letzten Zeugnis in Deutsch	[16] Mathe-matik	[17] Instrument-Nr. FB	[17] TH	[18] Teilnahmestatus Haupttest (TAVE) TH	[18] FB	[18] Nachtest (TAVE) TH	[18] FB	[19] Schüler-ID
001	Akcan, Deniz	2552019-001	2	nein	3	2	nein	nein	m	12.94	ja	ja	ja	m	12.94	3	2	1	1					2552019-001
002	Böning, Claudia	2552019-002	1	ja	2	2	nein	nein	w	03.95	ja	ja	ja	w	03.95	2	2	2	2					2552019-002
003	Freese, Verena	2552019-003	1	ja	5	3	nein	ja	w	08.95	nein	nein	ja					1	3					2552019-003
004	Frisch, Lukas	2552019-004	1	ja	4	4	nein	nein	m	11.94	nein	nein	nein					2	4					2552019-004
005	Günther, Marisa	2552019-005	1	ja	1	2	nein	nein	w	12.95	nein	ja	nein					1	1					2552019-005
006	Halberstadt, Julius	2552019-006	1	ja	2	3	nein	nein	m	02.95	nein	nein	nein					2	2					2552019-006
007	Liebig, Sebastian	2552019-007	1	ja	4	5	ja	ja	m	01.95	ja	ja	ja	m	01.95	4	5	1	3					2552019-007
008	Landmann, Anne-Kathrin	2552019-008	1	ja	3	1	nein	nein	w	02.95	ja	ja	ja	w	02.95	3	1	2	4					2552019-008
009	Marx, Alexander	2552019-009	1	ja	3	4	nein	nein	m	12.94	nein	nein	nein					1	1					2552019-009
010	Müller, Renée	2552019-010	2	ja	3	3	nein	nein	w	10.94	ja	ja	ja	w	10.94	3	3	2	2					2552019-010

Zum Verbleib in der Schule!

Versand im vorfrankierten und adressierten Umschlag an das DPC ohne Absenderinformation

Zur Aushändigung an den Testleiter!

A.4 "Lehrerliste"—List of teachers

Each school coordinator assigned teachers to classes using the list of teachers. Teachers could be assigned to different classes in different roles (e. g. class teacher in class A and math teacher in class B). Moreover, to one class there could be assigned several teachers, even in the same role. Depending on the way how German and math is organized (either in class or in courses) schools got different versions of this list as illustrated in figure 6 and figure 7.

NEPS-Haupterhebung K5/K9
Herbst/Winter 2010

Schulname:

Schul-ID:

Lehrerliste

Bitte listen Sie hier alle Klassen-, Deutsch- und Mathematiklehrkräfte, die in den NEPS-Klassen unterrichten.

NEPS-Haupterhebung K5/K9
Herbst/Winter 2010

Schulname:

Schul-ID:

Lehrerliste

Bitte kreuzen Sie hier an, welche Funktion die in Spalte [2] genannte Person in der jeweiligen Klasse (Spalten [4] bis [7]) innehat:
(Klassenlehrkraft = KL, Deutschlehrkraft = DL, Mathematiklehrkraft = ML)

[1]	[2]	[3]	[4]			[5]			[6]			[7]		
lfd. Nr. Lehrkraft	Vor- und Zuname der Lehrkraft	lfd. Nr. Lehrkraft	Klasse 5a			Klasse 5b			Klasse 9a			Klasse 9b		
			KL	DL	ML	KL	DL	ML	KL	DL	ML	KL	DL	ML
1		1												
2		2												
3		3												
4		4												
5		5												
6		6												
7		7												
8		8												
9		9												
10		10												

Achtung: Diesen Abschnitt nicht faxen!

Bitte senden Sie diesen rechten Abschnitt der ausgefüllten Lehrerliste bis spätestens XX.XX.2010 an das DPC, Fax-Nummer: 040 48500-501.

Figure 6: List of teachers for schools where math and german is teached in a class context

NEPS-Haupterhebung K5/K9
Herbst/Winter 2010

Schulname:

Schul-ID:

Lehrerliste

Bitte listen Sie hier alle Klassen-, Deutsch- und Mathematiklehrkräfte, die in den NEPS-Klassen unterrichten.

NEPS-Haupterhebung K5/K9
Herbst/Winter 2010

Schulname:

Schul-ID:

Lehrerliste

Bitte geben Sie hier an, welche Funktion die in Spalte [2] genannte Person in der jeweiligen Klasse (Spalten [4] bis [6]) innehat.

[1]	[2]	[3]	[4]				[5]	[6]
lfd. Nr. Lehrkraft	Vor- und Zuname der Lehrkraft	lfd. Nr. Lehrkraft	ist Klassenlehrkraft in				unterrichtet die Deutsch-kurse (lt. Kursliste)	unterrichtet die Mathematik-kurse (lt. Kursliste)
			(bitte ankreuzen)				*(bitte nur die lfd. Nr. aus der Kursliste eintragen)*	*(bitte nur die lfd. Nr. aus der Kursliste eintragen)*
			5a	5b	9a	9b		
1		1						
2		2						
3		3						
4		4						
5		5						
6		6						
7		7						
8		8						
9		9						
10		10						

Achtung: Diesen Abschnitt nicht faxen!

Bitte senden Sie diesen rechten Abschnitt der ausgefüllten Lehrerliste bis spätestens XX.XX.2010 an das DPC, Fax-Nummer: 040 48500-501.

Figure 7: List of teachers for schools where math and german is teached as courses

A.5 Modules of parental interview

Table 21: Modules of parental interview (CATI, 1st wave) by starting cohort

#	CATI modules	SC2	SC3	SC4
1	Contact module	x	x	x
2	Socio-demography of child	x	x	x
3	Siblings of child	x		
4	History of child care	x		
5	Early school enrolment	x		
6	Domestic activities of child	x		
7	Domestic learning environment	x		
8	Language Training	x		
9	Pre-schooling history		x	x
10	School history		x	x
11	Check module (X-module)		x	x
12	Schooling cross-section		x	x
13	Private Lessons		x	x
14	German classes		x	
15	Support		x	
16	Vocational-choice-support			x
17	Health of child	x	x	x
18	Strengths and Difficulties Questionnaire	x		x
19	Cultural capital	x	x	x
20	Socio-demography of interviewed parent	x	x	x
21	Social capital and segmented assimilation	x	x	x
22	Education of interviewed parent	x	x	x
23	Employment of interviewed parent	x	x	x
24	Partnership of interviewed parent	x	x	x
25	Socio-demography of partner of interviewed parent	x	x	x
26	Education of partner of interviewed parent	x	x	x
27	Employment of partner of interviewed parent	x	x	x
28	Residence	x	x	x
29	Household context	x	x	x
30	Household income	x	x	x
31	Wealth			x
32	Language competence and language usage	x	x	x
33	Identity, Orientation, and Transnationalism			x
34	Position generator			x
35	Role conceptions			x
36	Satisfaction with Kindergarten	x		
37	Satisfaction with school		x	x

Table 22: Datasets containing data from the parent interview by starting cohort (1st wave)

Data files	Unique ID	Content	SC
xParent	ID_p	Most data from CATI interview stored in cross-sectional format (1 row = 1 responding parent)	2 / 3 / 4
spChildCare	ID_p sptype sptypnr	History of child care for target child in spell format (1 row = 1 care spell)	2
spSchool	ID_p spell	History of schooling for target child in spell format (1 row = 1 school spell)	3 / 4
spGap	ID_p spell	Gap episodes in school history of target child in spell format (1 row = 1 gap spell)	3 / 4
ParentMethods	ID_p	Para-data in a cross-sectional format (e.g. date of interview, interview and interviewer characteristics, or response codes); contains information on all contacted parents not only on realized interviews, thus, contains more cases than xParents;	2 / 3 / 4

B Exemplary Items from test booklets in starting cohort 3 and 4

Beispielitems Hörverstehen auf Wortebene (PPVT)

Altersgruppe Klasse 9: Wort-Bild-Zuordnung

Instruktion

Wir beginnen in Zeile ‚A'. In dieser Zeile seht ihr 4 Bilder. Sucht das Bild, das zu ‚Löffel' passt. Die richtige Lösung ist das erste Bild. Kreuzt das entsprechende Kästchen in eurem Heft an.

Beispiel B: ‚krabbeln'.

Beispiel C: ‚Tau'.

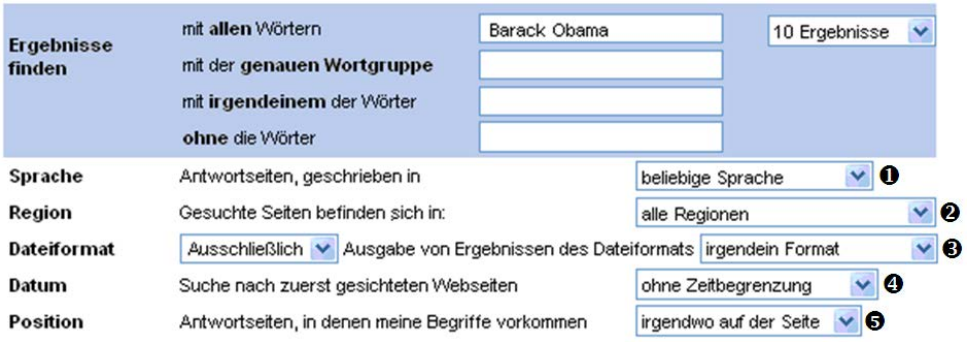
Beispiel D: ‚Gleichgewicht'.

Figure 8: Exemplary test item for listening in grade 9

Frage 8:

ITT915

Um Informationen über den aktuellen amerikanischen Präsidenten zu finden, hast du den Suchbegriff „Barack Obama“ eingegeben. Welche Auswahllisten würdest du benutzen, um Informationen in deutscher Sprache aus Deutschland zu bekommen?



Bitte kreuze in jeder Zeile die richtige Antwort an!

	richtig	falsch
❶ beliebige Sprache	☐	☐
❷ alle Regionen	☐	☐
❸ irgendein Format	☐	☐
❹ ohne Zeitbegrenzung	☐	☐
❺ irgendwo auf der Seite	☐	☐

Figure 9: Exemplary test item for ICT competence (1)

Frage 20: ITT943

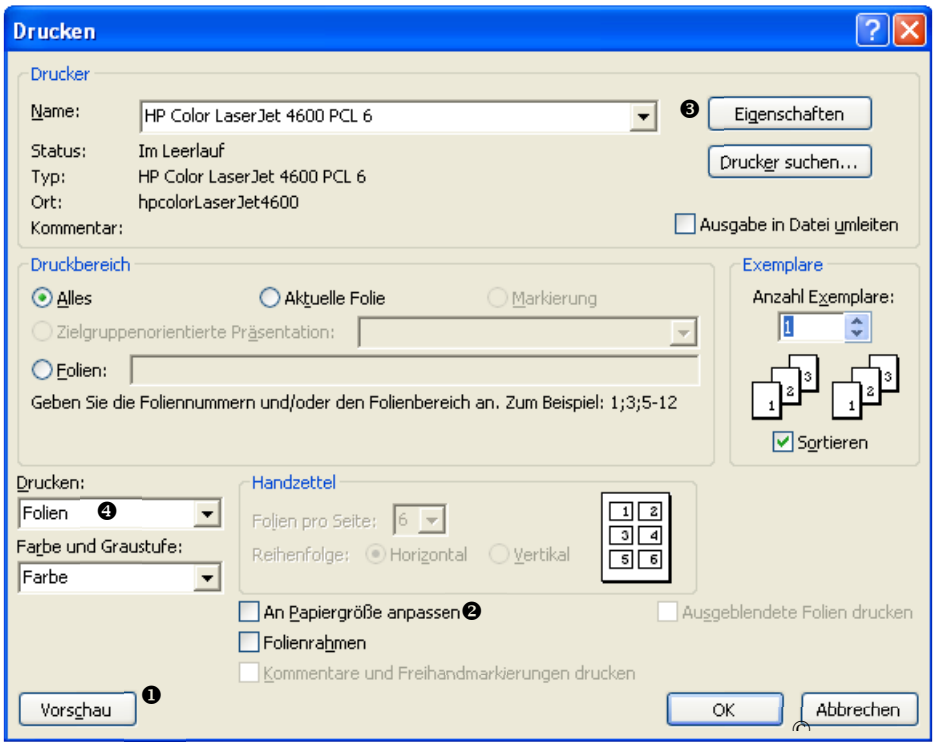
Rainer möchte Teile des Fotos mit seinem Bildbearbeitungsprogramm ausschneiden. Wohin muss er mit dem Mauszeiger klicken?

Bitte kreuze die richtige Antwort an! Bitte kreuze nur ein Kästchen an!

☐	❶ „Bearbeiten“
☐	❷ „Bild“
☐	❸
☐	❹

Figure 10: Exemplary test item for ICT competence (2)

Peter möchte seine Power-Point-Präsentation als Handzettel (Übersicht mit 6 Folien je Seite) ausdrucken. Wie geht er vor?



Bitte nur ein Kästchen ankreuzen ITT917

❶ Bei „Vorschau“ klicken	☐
❷ Bei „An Papiergröße anpassen“ klicken	☐
❸ Bei „Eigenschaften“ klicken	☐
❹ Bei „Folien“ klicken	☐

Figure 11: Exemplary test item for ICT competence (3)

Diese Tabelle wird dafür benutzt, um die Leihgebühren für jeden Schüler in der Musik-AG der Schule festzuhalten.

	A	B	C	D	E	F	G	H
1	**Name**	**Instrument**	**Monats-gebühr**	**Zahl der Monate**	**Gesamt-kosten**			
2	Christiane	Saxophon	45,00 €	10	450,00 €			
3	Katja	Klarinette	35,00 €	10	350,00 €			
4	Stefan	Trompete	60,00 €	5	300,00 €			
5	Ahmet	Trommeln	55,00 €		- €			
6	René	Flöte	42,00 €		- €			
7		**Gesamte Leihgebühren**			**1.100,00 €**			

Mit welcher Formel können die Gesamtkosten des ganzen Jahres für alle Instrumente berechnet werden?

Bitte nur ein Kästchen ankreuzen ITT956

=SUMME(D2:E6)	☐
=ADDIEREN(D2:E6)	☐
=SUMME(E2:E6)	☐
=ADDIEREN(E2:E6)	☐

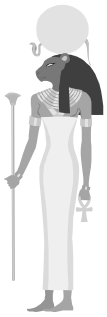
Figure 12: Exemplary test item for ICT competence (4)

ÄGYPTEN - PYRAMIDEN AM NIL

(1) Entlang des Nils zieht sich ein schmaler Streifen fruchtbaren Landes. Der Rest von Ägypten ist Wüste. Doch dieser grüne Streifen war genug, um dort eine der ersten großen Kulturen der Menschheit entstehen zu lassen: Das Alte Ägypten.

(2) Den ägyptischen Staat führten die Pharaonen. Sie waren aber nicht nur Könige, sie wurden von den Ägyptern sogar als Götter verehrt. Sie wurden verehrt und nach ihrem Tode höchst prunkvoll beerdigt. Die Ägypter glaubten an ein Leben nach dem Tode und sie glaubten, dass der tote Mensch seinen Körper noch einmal benötigen würde.

(3) In den berühmtesten Gräbern, den Pyramiden, wurden keine Mumien gefunden. Wahrscheinlich waren sie, mitsamt ihren Schätzen, Opfer von Grabräubern geworden.



(4) Die Pyramiden von Gizeh sind die berühmtesten Überbleibsel des Pharaonenreiches. Fast unglaublich: Denn die alten Ägypter hatten damals noch nicht einmal Werkzeuge aus Eisen, mit denen sie die Steinblöcke hätten bearbeiten können. Ihre Beile, Hämmer und Meißel waren aus Kupfer oder Stein.

(5) Ägypten war sehr lange Zeit das mächtigste und reichste Land der Erde. Es überstand zahlreiche Eroberungen, Kriege und Fremdherrschaften, verlor aber doch nach und nach an Macht und Einfluss.
Nachdem die letzte ägyptische Herrscherin Cleopatra 30 vor Christus, also 3.000 Jahre nach der ersten Pharaonendynastie, an einem Schlangenbiss gestorben war, machte der römische Kaiser Augustus Ägypten zu einem Teil seines Reiches.
Damit war die Geschichte des Alten Ägypten endgültig zu Ende.

Figure 13: Exemplary test item reading competence (1)

Nutze die Informationen über Ägypten auf der gegenüberliegenden Seite, um die unten stehenden Fragen zu beantworten.

Frage 1 (2.2)	**Die Kultur des Alten Ägyptens entstand entlang des Nils.** **Was war laut Text der Grund dafür?** **Bitte kreuze die richtige Antwort an! Bitte kreuze nur ein Kästchen an!**

- ☐ Man konnte so die Steine für die Pyramiden mit Booten transportieren.
- ☐ Das andere Land war von den Römern besetzt.
- ☐ Die Steinbrüche für den Bau der Pyramiden lagen in der Nähe.
- ☐ Der Rest des Landes war unfruchtbare Wüste.

Frage 3 (1.2)	**Im Text erfährt man, auf welche Weise Cleopatra gestorben ist.** **Was war die Ursache ihres Todes?** **Bitte kreuze die richtige Antwort an! Bitte kreuze nur ein Kästchen an!**

- ☐ Sie starb durch die Hand Julius Cäsars.
- ☐ Cleopatras Leben endete aufgrund ihres hohen Alters.
- ☐ Sie wurde von einer giftigen Schlange gebissen.
- ☐ In einem Krieg verlor Cleopatra das Leben.

Figure 14: Exemplary test item reading competence (2)

Frage 7 (3.1)

Welche Teilüberschrift passt am besten zu welchem Textabschnitt?

Ordne die Buchstaben von A bis F den Textabschnitten zu!

Trage die passenden Buchstaben in die Tabelle ein! Ein Buchstabe bleibt übrig.

Abschnitt	Teilüberschrift
1	☐
2	☐
3	☐
4	☐
5	☐

A Fruchtbares Land

B Gräber und Schätze

C Werkzeuge für Pyramiden

D Wichtige Mumienfunde

E Cleopatra

F Götter und Könige

Figure 15: Exemplary test item reading competence (3)

Sample Item 1: **Stones in the Bowl**		
Age cohort(s)	*Content area*	*Cognitive component(s)*
kindergarten – grade 1	quantity (sets, numbers & operations)	mathematical problem solving, technical abilities and skills

There are four stones in this bowl. Now I add three more stones. [The bowl is covered so that the child cannot look inside.] Can you tell me, how many stones there are in the bowl now?

Sample Item 2: **The Fence**		
Age cohort(s)	*Content area*	*Cognitive component(s)*
Grade 5 – grade 7	Space & Shape	problem solving, technical abilities and skills

Mr. Brown owns a rectangular piece of land and wants to fence it in. He has already made some calculations and then bought a 40 *m* fence. The piece of land has a width of 8 *m*. How long is the land?

☐	5 *m*
☐	8 *m*
☐	12 *m*
☐	16 *m*

Figure 16: Exemplary test item mathematical competence (1)

Sample Item 3: **At the Zoo**

Age cohort(s)	*Content area*	*Cognitive component(s)*
Grade 9 – College students	Change & Relationships	using representational forms, communication

In summertime, the Atown-Zoo is visited by more people than in wintertime. Yet this year, a brown bear was born in October, which attracted a lot of visitors due to its cuteness. Which of the following graphs does correctly display the number of visitors at the Atown-Zoo?

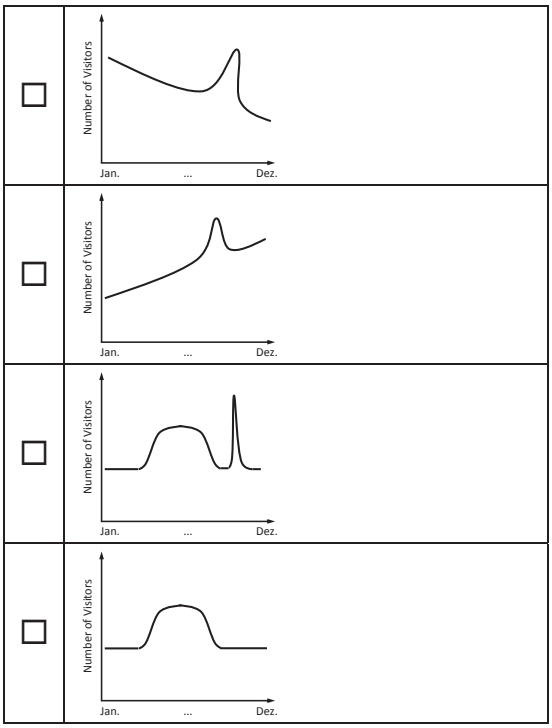


Figure 17: Exemplary test item mathematical competence (2)

HERZ-KREISLAUF-SYSTEM

Kaffee wirkt aufgrund des Koffeingehaltes anregend auf den Kreislauf. Menschen, die bereits unter hohem Blutdruck leiden, sollten allerdings soweit wie möglich auf den Genuss von Kaffee verzichten und auf koffeinfreie Getränke ausweichen. Wissenschaftler vermuten nun aber, dass auch koffeinfreier Kaffee den Blutdruck eines Menschen erhöht.

Wie könnte eine entsprechende Studie angelegt sein?

	Testgruppe	Messung	Zeitpunkt der Messung
Versuchsplan A	Eine Testgruppe: trinkt zunächst koffeinfreien Kaffee, dann koffeinhaltigen Kaffee	Blutdruck	Vor und nach dem Trinken
Versuchsplan B	Gruppe 1 trinkt koffeinfreien Kaffee, Gruppe 2 trinkt koffeinhaltigen Kaffee	Blutdruck	Nach dem Trinken
Versuchsplan C	Gruppe 1 trinkt koffeinfreien Kaffee, Gruppe 2 trinkt Wasser	Blutdruck	Nach dem Trinken
Versuchsplan D	Gruppe 1 hat Bluthochdruck, Gruppe 2 ist gesund. Beide Gruppen trinken koffeinfreien Kaffee	Blutdruck	Vor und nach dem Trinken

Bitte kreuzen Sie die richtige Antwort an! Bitte kreuzen Sie nur ein Kästchen an!

0	Versuchsplan A
0	Versuchsplan B
0	Versuchsplan C
0	Versuchsplan D

Figure 18: Exemplary test item science competence

Beispielaufgabe Klasse 6 /Klasse 9:

Peter hat diese Woche viel zu tun: Er soll zwei Mal im Verein schwimmen, hat viele Hausaufgaben und muss ein Geburtstagsgeschenk für seinen Freund besorgen.

Wie sollte er vorgehen, um alles zu schaffen?

Bitte beurteile die Nützlichkeit aller Vorschläge.

	gar nicht nützlich	kaum nützlich	etwas nützlich	sehr nützlich
Er macht einen Plan für die Woche und teilt sich die Zeit für die Aufgaben ein. Seinen Plan hält er sehr genau ein.	☐	☐	☐	☐
Er verbindet verschiedene Aufgaben miteinander. Das Geschenk kauft er auf dem Weg zum Schwimmverein.	☐	☐	☐	☐
Er lässt sich von anderen helfen. Er bittet seinen Bruder, das Geschenk zu kaufen.	☐	☐	☐	☐
Er macht nur die Hausaufgaben, die schnell gehen. Dann macht er die anderen Dinge.	☐	☐	☐	☐
Er besorgt zuerst das Geburtstagsgeschenk. Wenn das zu lange dauert, lässt er die Haus-aufgaben oder das Schwimmen weg.	☐	☐	☐	☐
Er macht immer genau die Dinge, auf die er gerade am meisten Lust hat.	☐	☐	☐	☐

Figure 19: Exemplary item for declarative metacognition

B References

Allmendinger, Jutta et al. (2011). "Adult education and lifelong learning". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, 283–299. DOI: 10.1007/s11618-011-0178-3.

Aschinger, Florian et al. (2011). "Higher education and the transition to work". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, 267–282. DOI: 10.1007/s11618-011-0190-7.

Aßmann, Christian, Hans Walter Steinhauer, Hans Kiesl, et al. (2011). "Sampling designs of the National Educational Panel Study: Challenges and Solutions". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 51–65. DOI: 10.1007/s11618-011-0181-8.

Aßmann, Christian, Hans Walter Steinhauer, and Sabine Zinn (2012). *Weighting the Fifth and Ninth Grader Cohort Samples of the National Educational Panel Study, Panel Cohorts.* Tech. rep. German National Educational Panel Study (NEPS).

Aust, Folkert, Doris Hess, and Angela Prussog-Wagner (2013, forthcoming[a]). *Methodenbericht - NEPS Elternbefragung von Kindergartenkindern, Startkohorte 2 - Kindergarten, Haupterhebung 1. Welle 2011 (Teilstudie B11).* Tech. rep. infas Institut für angewandte Sozialwissenschaft GmbH.

– (2013, forthcoming[b]). *Methodenbericht: NEPS Elternbefragung von Schülern (Startkohorte 4 - Klasse 9), Haupterhebung 1. Welle 2011 (Teilstudie B34).* Tech. rep. Bonn: infas Institut für angewandte Sozialwissenschaft GmbH.

Berendes, Karin et al. (2011). "Kindergarten and Elementary School". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 203–216.

Blossfeld, Hans-Peter (1985). *Bildungsexpansion und Berufschancen.* Frankfurt: Campus.

Blossfeld, Hans-Peter, Jutta von Maurice, and Thorsten Schneider (2011). "The National Educational Panel Study: need, main features, and research potential". In: ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice. Special Issue 14. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 5–17. DOI: 10.1007/s11618-011-0178-3.

Blossfeld, Hans-Peter, Hans Günther Roßbach, and Jutta von Maurice, eds. (2011). *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).*

Edele, Aileen, Kristin Schotte, Martin Hecht, et al. (2012). *Listening comprehension tests of immigrant students' first language (L1) Russian and Turkish in grade 9: Scaling procedure and results.* NEPS Working Paper 13. Bamberg.

Edele, Aileen, Kristin Schotte, and Petra Stanat (forthcoming). *Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Construction and validation.* Tech. rep. German National Educational Panel Study (NEPS).

Erikson, Robert, John H. Goldthorpe, and Lucienne Portocarero (1979). "Intergenerational Class Mobility in Three Western European Societies. England, France and Sweden". In: *British Journal of Sociology* 30, pp. 415–441.

Frahm, Sarah et al. (2011). "Transition and development from lower secondary to upper secondary school". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 283–299. DOI: 10.1007/s11618-011-0188-1.

Ganzeboom, Harry B. G. (2010). *Questions and Answers about ISEI-08.* URL: http://home.fsw.vu.nl/hbg.ganzeboom/isco08/qa-isei-08.htm (visited on 10/09/2012).

Ganzeboom, Harry B. G. et al. (1992). "A standard international socio-economic index of occupational status". In: *Social Science Research* 21, pp. 1–56.

German National Educational Panel Study (NEPS) (2013). *Starting Cohort 4: 9th Grade (SC4), Waves 1 and 2, Questionnaires (SUF Version 1.1.0).* Tech. rep. Bamberg: German National Educational Panel Study (NEPS).

Jann, Ben (2007). *fre: Stata module to display one-way frequency table.* URL: http://ideas.repec.org/c/boc/bocode/s456835.html (visited on 10/09/2012).

Koberg, Tobias (2012a). *Starting Cohort 4 of the National Educational Panel Study: Anonymisation procedures and statistical disclosure control.* Tech. rep. German National Educational Panel Study (NEPS).

– (2012b). *Starting Cohorts 2, 3, 4, and 6. Data Manual [Supplement]. RegioInfas (infas geodaten). v 1.0 September 21, 2012.* Bamberg.

Lockl, Kathrin (2012a). *Assessment of declarative metacognition: Starting Cohort 4 – Ninth Grade.* Tech. rep. German National Educational Panel Study (NEPS).

– (2012b). *Assessment of procedural metacognition: Scientific Use File 2012, Version 1.0.0.* Tech. rep. German National Educational Panel Study (NEPS).

Ludwig-Mayerhofer, Wolfgang et al. (2011). "Vocational education and training and transitions into the labour market". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 251–266.

Lüttinger, Paul and Wolfgang König (1998). "Die Entwicklung einer international vergleichbaren Klassifikation für Bildungssysteme". In: *ZUMA Nachrichten* 22, pp. 1–14.

Pohl, Steffi and Claus H. Carstensen (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests*. NEPS Working Paper 14. Bamberg.

Pohl, Steffi, Kerstin Haberkorn, et al. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 4 in Ninth Grade*. NEPS Working Paper 16.

Rohwer, Götz (2011). *Using Sampling Weights for Model Estimation?* URL: http://steinhaus.stat.ruhr-uni-bochum.de/papers/dsw.pdf (visited on 10/09/2012).

Schimpl-Neimanns, Bernhard (2003). *Mikrodaten-Tools: Umsetzung der Berufsklassifikation von Blossfeld auf die Mikrozensen 1973–1998*. ZUMA-Methodenbericht 2003/10.

Schlesiger, Claudia et al. (2011). "From birth to early child care". In: ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice. Special Issue 14. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 187–202. DOI: 10.1007/s11618-011-0178-3.

Senkbeil, Martin and Jan-Marten Ihme (2012). *NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 4 in Ninth Grade*. Tech. rep. 17. German National Educational Panel Study (NEPS).

Skopek, Jan, Sebastian Pink, and Daniel Bela (2012). *Starting Cohort 3: 5th Grade (SC3). SUF Version 1.0.0. Data Manual*. Bamberg.

– (2013). *Starting Cohort 4: 9th Grade (SC4). SUF Version 1.1.0. Data Manual*. NEPS Research Data Paper. National Educational Panel Study (NEPS), University of Bamberg.

Treiman, Donald J. (1977). *Occupational prestige in comparative perspective*. New York et al: Academic Press.

UNESCO (2006). *ISCED 1997. International Standard Classification of Education*. United Nations Educational, Scientific and Cultural Organization. URL: http://www.uis.unesco.org/Library/Documents/isced97-en.pdf.

Wagner, Wolfgang et al. (2011). "Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)*. Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 233–249.

Wegener, Bernd (1985). "Gibt es Sozialprestige?" In: *Zeitschrift für Soziologie* 14.3, pp. 209–235.

Weinert, Susanne et al. (2011). "Development of competencies across the life span". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)*. Ed. by Hans-Peter Blossfeld, Hans Günther Roßbach, and Jutta von Maurice, pp. 67–86.

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

– (2012b). *Starting Cohort 4: 9th Grade (SC4), SUF Version 1.0.0, Data Manual [Supplement]: Codebook (en)*. Tech. rep. German National Educational Panel Study (NEPS).

– (2012c). *Startkohorte 4: Klasse 9 (SC4), SUF-Version 1.0.0, Data Manual [Supplement]: Codebook (de)*. Tech. rep. German National Educational Panel Study (NEPS).