

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

SPONSORED BY THE



Federal Ministry
of Education
and Research

Copyrighted Material

University of Bamberg, National Educational Panel Study (NEPS), 96045 Bamberg

<https://www.neps-data.de>

Principal Investigator: Prof. Dr. Hans-Günther Roßbach

Vice Managing Director: Prof. Dr. Sabine Weinert

Executive Director of Research: Dr. Jutta von Maurice

Executive Director of Administration: Dipl. sc. pol. Univ. Dipl.-Betriebswirt (FH) Gerd Bolz
Bamberg, 2012

Data Manual

Starting Cohort 3

From Lower to Upper Secondary School

(NEPS SC3 1.0.0)

Jan Skopek, Sebastian Pink, Daniel Bela
NEPS Data Center

September 27, 2012

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).

Citation of the manual:

Skopek, J. S. Pink and D. Bela. (2012). Data Manual. Starting Cohort 3 – From Lower to Upper Secondary School. NEPS SC3 1.0.0. NEPS Research Data Paper, University of Bamberg.

This release of scientific use data from Starting Cohort 3 – “From Lower to Upper Secondary School” 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 preparing and editing

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)

Dietmar Angerer (filtering syntax development)

Maja-Henrieke Lomb (data plausibility testing)

Judith Pfohl (data plausibility testing)

Frauke Erzigkeit (proofreading)

Data manual

Jan Skopek

Sebastian Pink

Daniel Bela

We would like to thank several researchers from the NEPS network who participated in testing a pre-version of these data. Their feedback contributed to a significant improvement of the scientific use file for this main release.

National Educational Panel Study (NEPS)

Data Center

Wilhelmsplatz 3

96047 Bamberg, Germany

Contact: userservice.neps@uni-bamberg.de

Web: <https://www.neps-data.de/de-de/datenzentrum>

Phone: +49 951 8633511

Table of Contents

| | | |
|-----|--|----|
| 1 | Introduction | 3 |
| 1.1 | About this manual | 3 |
| 1.2 | Obtaining the data | 4 |
| 1.3 | Three modes of data access | 4 |
| 1.4 | Respect rules of data usage! | 5 |
| 1.5 | Publications with NEPS data | 5 |
| 2 | General Conventions | 6 |
| 2.1 | File names | 6 |
| 2.2 | Variable names | 7 |
| 2.3 | Missing values | 2 |
| 3 | Surveys and Sampling | 4 |
| 3.1 | Overview | 4 |
| 3.2 | Sampling | 5 |
| 3.3 | Surveys and Tests | 6 |
| 4 | Data Structure | 2 |
| 4.1 | Overview | 2 |
| 4.2 | Data files | 3 |
| 4.3 | Syntax for cleaning filtered question data (PAPI) | 9 |
| 5 | Coding | 9 |
| 6 | Weights | 16 |
| 7 | Examples | 16 |
| 7.1 | Example 1 – First steps using CohortProfile | 17 |
| 7.2 | Example 2 – Merge datasets to CohortProfile | 18 |
| 7.3 | Example 3 – Indirectly matching different datasets | 19 |
| 7.4 | Example 4 – Prepare multilevel data | 21 |
| 7.5 | Example 5 – Using cleaning syntax for PAPI filtering | 23 |
| 7.6 | Example 6 – Using weights | 24 |
| 8 | Rules and Recommendations | 25 |
| 8.1 | Rules | 25 |
| 8.2 | Recommendations | 25 |
| 9 | Tools for Stata users | 26 |
| 10 | Further information | 28 |

| | | |
|------|---|----|
| 11 | Appendix | 29 |
| 11.1 | Characteristics of the survey of students with special educational needs..... | 29 |
| 11.2 | “Schülerliste” – List of students | 31 |
| 11.3 | “Lehrerliste” – List of teachers | 34 |
| 11.4 | Modules of parental interview | 36 |
| 11.5 | Exemplary Items from test booklets in starting cohort 3 and 4 | 38 |
| 12 | References | 50 |

1 Introduction

1.1 About this manual

This manual is intended to assist your work with the data of the *NEPS Starting Cohort 3 – From Lower to Upper Secondary School* (NEPS SC3 1.0.0). 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 comprehensive documentation resource. Please consult our website

<https://www.neps-data.de/de-de/datenzentrum> (in German)

<https://www.neps-data.de/en-us/datacenter> (in English)

for background information on the studies, survey instruments, a structured documentation, and many more resources.

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

Table 1 Documentation supplements for starting cohort 3

| Description | Language | Status |
|---|----------|-------------|
| Codebook | DE/EN | forthcoming |
| Survey instruments package | DE/EN | forthcoming |
| Technical report: weighting | EN | forthcoming |
| Technical report: competence scaling | EN | forthcoming |
| Technical report: Anonymisation | EN | forthcoming |
| Technical report: RegioInfas / infas geodaten | EN | forthcoming |
| Technical report: Data edition | EN | forthcoming |
| Field work reports from infas | DE | forthcoming |
| Field work reports from IAE-DPC | DE | forthcoming |
| Syntax package (e.g. cleaning syntax for PAPI data) | Stata | forthcoming |

You can download these documents here:

<https://www.neps-data.de/de-de/datenzentrum/forschungsdaten/startkohorteklasse5>
(in German)

<https://www.neps-data.de/en-us/datacenter/researchdata/startingcohort5thgrade>
(in English)

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

1.2 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, click here for instructions:
 - For German users:
<https://www.neps-data.de/de-de/datenzentrum/datenzugangsweg/datennutzungsverträge>
 - For non-German users:
<https://www.neps-data.de/en-us/datacenter/dataaccess/datauseagreements>
- After approval, sign in as a registered NEPS user at the login at 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.

1.3 Three modes of data access

NEPS offers to you three modes of access to the data:

- Download from our website www.neps-data.de ,
- 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 authentication 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 anonymisation. 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:

<https://www.neps-data.de/de-de/datenzentrum/datenzugangsweg>
(in German)

<https://www.neps-data.de/en-us/datacenter/dataaccess>
(in English)

This concept of data dissemination translates into an “onion-shaped” model of datasets: The most sensitive data (“on-site”) that include weakly anonymised information in high resolution represent the outer layer, with “remote access” and “download” levels being

subsets of these data. Thus, 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 Supplement “Anonymisation” (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). SPSS files are delivered separately in both languages.

1.4 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! Read more in section 8.1.

1.5 Publications with NEPS data

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

Blossfeld, H.-P., H.-G. Roßbach, and J. von Maurice (eds.) (2011). “Education as a Lifelong Process – The German National Educational Panel Study (NEPS)”, Zeitschrift für Erziehungswissenschaft: Special Issue 14.

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 3 – 5th grade (From Lower to Upper Secondary School), doi:10.5157/NEPS:SC3:1.0.0. 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. The DOI of this release redirects to a landing page providing basic information on the data:

<http://dx.doi.org/10.5157/NEPS:SC3:1.0.0>

2 General Conventions

2.1 File names

The names of the datasets included in this release were defined by a number of conventions which are displayed in Table 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 <i>xTarget</i> contains cross-sectional data from the target questionnaire; spSchool contains schooling spells;) Filenames of generated datasets do not have a prefix and always start with a capital letter (e.g., <i>CohortProfile</i> , <i>Biography</i>) |
| [D,R,O] | Confidentiality Level D = Download version R = Remote access version O = On-site 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; 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-suffix: 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. |

To give an example, the physical file “SC3_xTarget_D_1-0-0.dta” refers to the download-version for the data file xTarget of starting cohort 3 of data release 1-0-0.

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.

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 the following example (see Figure 1). More detailed information is given in Table 3.

Figure 1 Example for variable naming

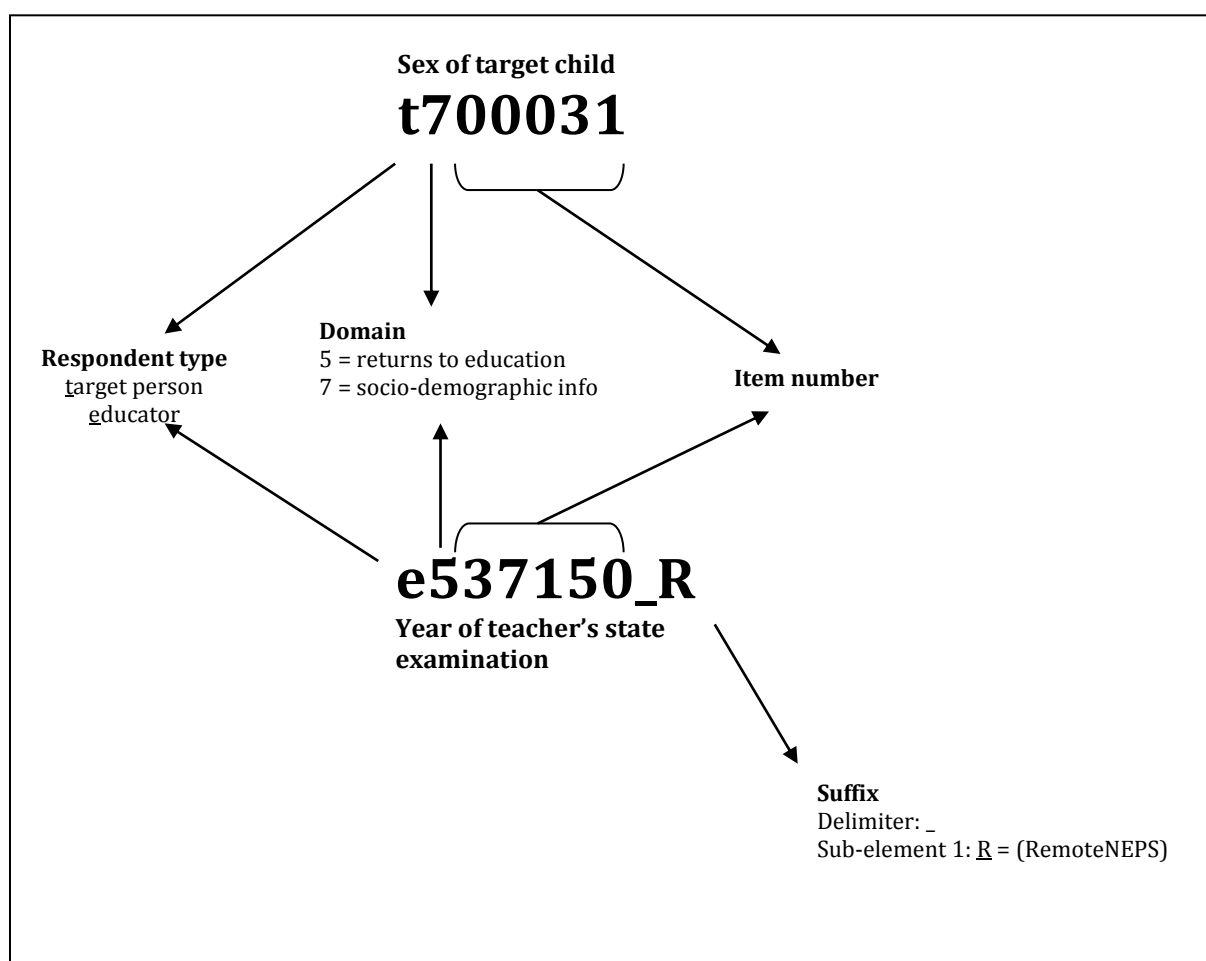


Table 3 Naming conventions of variable names

| Digit | Description |
|-------|--|
| 1 | Indicates to which respondent type the variable refers; in starting cohort 3, this character can be <i>t</i> (target person), <i>p</i> (one parent of target person), <i>e</i> (educator) and <i>h</i> (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). |
| 2 | <p>Topic/domain (according to the theoretically coordinated dimensions of the NEPS):</p> <ul style="list-style-type: none"> 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 | <p>Suffix (optional): Suffixes are separated from the previous characters by an underscore. There are four types of suffixes:</p> <ul style="list-style-type: none"> • Suffixes for generated variables: Generated variables are indicated by the suffix <i>_g#</i> (<i>_g1</i>, <i>_g2</i>, etc.). In most cases, the running number after <i>_g</i> is a simple enumerator. However, there are two types of generated variables that assign meanings to these running numbers: regional and occupational variables. <ul style="list-style-type: none"> ○ Regional codes based on the Nomenclature of Territorial Units for Statistics (NUTS) <ul style="list-style-type: none"> ▪ <i>g1</i> = NUTS level 1 (federal state/Bundesland) ▪ <i>g2</i> = NUTS level 2 (government region/Regierungsbezirk) ▪ <i>g3</i> = NUTS level 3 (district/Kreis) ○ Occupational/prestige codes <ul style="list-style-type: none"> ▪ <i>g1</i>: KldB 1988 (German Classification of Occupations 1988) ▪ <i>g2</i>: KldB 2010 (German Classification of Occupations 2010) ▪ <i>g3</i>: ISCO-88 (Internat. Standard Classification of Occupations 1988) ▪ <i>g4</i>: 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)
 - 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 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 anonymised (see 1.4). The suffix indicates a variable's degree of anonymisation. This suffix may either standalone (e.g., country of birth: *t405010_R*) or be combined with other suffixes (e.g., district of place of birth: *t700101_g3R*)
 - 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 getting first measurement there), and the third part defines the item number. Table 4 gives an overview to the logic of parts. There are two versions of item variables: scored items named *{varname}_c* and scored partial credit-Items named *{varname}_s_c*. Moreover, suffix *_sc{number}* is used for several scores and the meaning of the suffixed number is fixed as follows: 1=WLE, 2=standard error of WLE, 3=sum, 4=mean, 5=difference. If there are several versions e.g. of a sum score letters are appended additionally (e.g. dgg9_sc3a is the first sum of DGCF). For example, variable *grk1_sc3* represents the sum score of the grammar test of children being tested for the first time in Kindergarten. To give another example, variable *vok10007_c* is a scored version (values 0 or 1) of a test item measuring vocabulary at children being tested for the first time in Kindergarten.

Table 4 Different parts of names of variables in test data

| Part I | | | Part II | | | Part III |
|------------|---------|-------------------------|--------------|---------|--|-------------|
| Instrument | Meaning | | Target Group | Meaning | | Item number |
| 2 chars | | | 2 chars | | | 4 chars |
| R | E | Read | N | 0 | Newborn 0 | |
| M | A | Math | ... | ... | ... | |
| S | C | Science | N | 3 | Newborn 3 | |
| I | C | ICT | K | 1 | KiGa 1 | |
| L | I | Listening | K | 2 | KiGa 2 | |
| V | O | Vocabulary | G | 1 | Grade 1 | |
| O | R | Orthography | G | 5 | Grade 5 | |
| G | R | Grammar | G | 9 | Grade 9 | |
| D | G | DGCF | G | A | Grade 10 | |
| E | F | English Foreign | G | B | Grade 11 | |
| M | P | Meta procedural | G | C | Grade 12 | |
| M | D | Meta declarative | G | D | Grade 13 | |
| R | S | Reading Speed | V | 1 | Vocation 1 | |
| A | T | Attention | ... | ... | ... | |
| N | R | Native Language Russian | V | 3 | Vocation 3 | |
| N | T | Native language Turkish | S | 1 | University students 1 | |
| V | I | Verbal Intelligence | ... | ... | ... | |
| N | I | Nonverbal Intelligence | S | 5 | University students 5 | |
| F | A | FAIR | A | 1 | Adults 1: 2009 | |
| | | | ... | ... | ... | |
| | | | A | 4 | Adults 4 | |
| | | | C | I | Cohort invariant (for instruments administered unchanged in all cohorts) | |

2.3 Missing values

We provide different missing codes for different situations 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 *nepsmis* 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 | Implausible value |
| -94 | Not reached (only applicable for competence tests) |
| -90 | Unspecific missing |
| -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 |
| -96 | Not in list |
| -99 | Filtered (in PAPI mode) |
| . | Filtered / system missing (in CATI/CAPI mode) |
| Edition missings (recoded into missing) | |
| -52 | Implausible value removed |
| -53 | Anonymised |
| -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).
 - Implausible 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 (-20, ... ,-29) pertain to variable-specific nonresponse categories (e.g. variable *p731801*, indicating educational achievements, has a missing code -20 for "no school-leaving qualification").

- Missings that occur for unknown reasons are coded by -90; this especially happens in PAPI questionnaires, where the cause for a respondent not answering a question cannot be determined.
- *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 also used 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).
- *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 anonymised (-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 in a questionnaire although it was administered, the concerning variables are assigned the missing code “Not participated” (-56). This missing code is special because target persons lacking interview data (e.g. due to illness) are usually 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 and update the *nepsmiss* command.

3 Surveys and Sampling

3.1 Overview

This data release comprises data from the first wave data of the NEPS fifth graders cohort, the NEPS starting cohort 3. In general, for starting cohort 3 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 5th grade in German schools in fall 2010. As a consequence, data collection at first wave was organized in two different NEPS studies: study A28/A63¹ covering regular schools and study A56 covering special needs schools. Moreover, field work procedures for the special needs sample deviated slightly from those of the regular sample, which is the bigger one. Hence, when analyzing the data you should always be aware of this fact.² Eventually, students were sampled in a multistage approach using an explicit stratification by form of school.

In the cohort of fifth graders, NEPS has initiated a class-based sampling design, which means that we sampled and followed up entire school classes. The fifth graders sample is surveyed annually. One of the major goals is capturing the children’s’ transition and development from lower secondary to upper secondary school. For an in depth treatment of theoretical perspectives and design issues of the school sample consult Frahm et al. (2011).

Students (target persons of this cohort) were tested by paper-based competence tests and surveyed by questionnaires in schools. The school survey has been organized as group testing. In general the testing time took one school day. Competence tests were administered at first. Afterwards students filled out a paper-and-pencil (PAPI) questionnaire (the students’ questionnaire).

Teachers, principals of schools (“Schulleiter”), and parents of children were surveyed as well. In wave 1 of starting cohort 3, class teachers (“Klassenlehrer”) and German teachers (“Deutschlehrer”) were surveyed. Surveying of math teachers (“Mathematiklehrer”) is projected in second wave of starting cohort 3. Besides a general teacher questionnaire with more general questions there were also special questionnaires for the class teachers and German teachers with specific questions 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 9th graders cohort (starting cohorts 2 and 4), a 30-minute-interview with parents has been conducted via computer assisted telephone

¹ The study A63 was originally intended as an oversample of migrants, which was later integrated in the main sample due to small case numbers.

² In general, we recommend to analyse those samples separately.

interview (CATI) to collect comprehensive context and background data on the target persons. This was organized in a separate study (B20) with almost parallel field time.

Participation in the NEPS study is voluntary. Parents of children were invited to participate in the study. Note that, contrary to starting cohort 2, in starting cohort 3 and 4 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 collection of wave 1 started in fall 2010. The panel sample comprises 6,112 fifth graders. Among these, 5,525 fifth graders attended 234 regular schools (study A28), and 587 fifth graders with special educational needs in the area of learning attended 57 special needs schools (study A56). All in all, 5,778 fifth graders participated in the first wave (therefore, questionnaire data and/or test data is available for this group). 334 students are temporary dropouts since they did not appear at the test days in schools (e.g. due to illness). Context data is available from 653 teachers and 205 principals. From of the set of contacted parents about 87.3% did participate in the telephone survey, about 10.3% temporary dropped out (i.e. will be contacted again in subsequent parent study), and about 2.4% finally dropped out (e.g. withdrew from survey participation).

The questionnaire and test items have been developed by NEPS pillar 1 (Competence Development), pillar 2 (Learning Environments), 3 (Educational Decisions), 4 (Migration background) and 5 (Returns to Education). NEPS stage 4 (From Lower to Upper Secondary School)—headed by Wilfried Bos (Institute for School Development Research at the Technical University of Dortmund) in collaboration with Inge Blatt (University of Hamburg)—provided stage specific items and a stage specific orthography test. Furthermore, NEPS stage 4 was responsible for the compilation of questionnaires.

Field work of A28/A63 (regular students) and A56 (special needs students) was organized and conducted by the data collection institute IEA DPC (IEA Data Processing and Research Center, Hamburg). Thereby, paper-based questionnaires and tests were administered to students, and further questionnaires to teachers and principals of the sampled schools (PAPI mode). Parental interviews (NEPS study B20) were accomplished via telephone interviews (CATI mode) from January to July 2011. All parental interviews have been conducted by the data collection institute infas (Institute for Applied Social Sciences, Bonn). Scientific use files are prepared and provided by the NEPS Data Center.

Table 6 provides an overview to respondents, modes, instruments, and contents of data collection of the first wave. In the following sections further details on sampling and survey design are given.

3.2 Sampling

Target persons are all children 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. Children 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, children not able to follow the normal testing procedure attending regular schools are excluded. Additionally, NEPS comprises a sample of children attending special needs schools. Here the focus is put on schools with main emphasis on learning disabilities. Access to this population is gained via special needs schools with federal-state-specific provisions explicitly for children with learning disabilities. Overall, 80% of children attending special needs schools have a diagnosed learning disability constituting the largest group of children 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. 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 children. 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 school types have been defined as explicit strata to sample schools.

For detailed discussion of the sampling design have a look at Aßmann et al. (2011). For details on sampling and weights consult Supplement “Weighting” (see section 1.1).

3.3 Surveys and Tests

3.3.1 Competence testing

Regular school students of starting cohort 3 were tested in different domains. In first wave in 2011, they were tested in domain-general cognitive functions, reading, mathematics and orthography as well as procedural metacognition. The test situation was organized as follows: group testing, normally taking place in the classroom, using individual tests; students are sitting in single seats, and are accompanied by one test instructor, plus normally one supervisor.

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

- **Test order test booklet 1**
reading speed, reading competence + procedural metacognition, mathematical competence + procedural metacognition, cognitive basic skills: perceptual speed + reasoning, orthography + procedural metacognition
- **Test order test booklet 2**
reading speed, mathematical competence + procedural metacognition, reading competence + procedural metacognition, cognitive basic skills: perceptual speed + reasoning, orthography + procedural metacognition

In general, testing time took one school day and was organized in groups. There was one 15-minute break before the test of cognitive basic skills. The duration of the test was 100.5 minutes.

Learn more about the test situation of 5th grade students in regular schools by consulting the document “Main Study 2010/11 (A28), Students, 5th Grade, Regular Schools –Information on the Competence Test” being part of Supplement “Survey instruments package” (see section 1.1). Additionally, consult the technical report on competence scaling (Supplement). Moreover, Weinert et al. (2011) provides a comprehensive discussion on the measurement of competencies across the life span in the NEPS.

Note that competence data is available only for regular students. Due to validity concerns test data of the special needs student sample is not included in this release. For further information, please contact pillar 1 of the NEPS³, which accounts for editing, scoring, and releasing of competence data.

3.3.2 Data from students

After testing, students received a student’s questionnaire. It covers several topics like sociodemographics, learning environments, cultural capital, educational aspirations, satisfaction, health, language use, reading engagement, or quality of instruction.

Naturally, collecting questionnaire data from students with special educational needs (“Förderschüler”) is challenging. Therefore, NEPS developed different field work procedures and instruments for students from the special needs sample. For example, there were several rotations for the student’s questionnaire with different arrangements of topics (see section 11.1, p. 29ff). Moreover, questions were read aloud by the supervisor.

3.3.3 Data on students and home context provided by parents

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 *B20* for starting cohort 3). Parents of children have been interviewed via telephone (CATI mode) based on their consent that was given in advance. Preferably selected for interview were such 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 has been delivered to all parents of the three starting cohorts. However, there are question blocks that were delivered specifically by cohort, e.g. only to starting cohort 2 (like the child care history).

³Web: <https://www.neps-data.de/en-us/projectoverview/structure/pillars/competencedevelopment.aspx>

Table 15 in the appendix (p. 36) 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 (Supplement).

3.3.4 Contextual data on classes and schools

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.

There were separate questionnaires for the class teacher and the German teacher. The first mainly contained questions on structural characteristics and composition of the class. The latter asked specific questions regarding the quality of instruction of the German class.

Similarly, 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.

3.3.5 Data from lists of students and lists 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 11.2 for an example of the "Schülerliste" and section 11.3 for examples of the "Lehrerliste".

In the first list all students—whether participating or not—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) a subset of the basic data on student is provided in the datasets from this list; regardless whether students or parents responded to questionnaires.

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). E.g. there is about 99% agreement regarding sex and about 98% agreement regarding year of birth (very similar to the accordance between student and parent responses). However,

when working with these variables one should remember the fact that variables looking similar may be based on different data-generation processes (data provided by coordinators, students, and parents).

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, and in which function (class, German, or math teacher, see Figure 8 on page 34). Thus, a multiple assignment was possible. For such schools where German and math is organized in form of “courses” (Kurse) instead of classes, the coordinator could define one or more courses per teacher (see Figure 9 on page 35).

3.3.6 Logic of assignment of students and teachers to classes

As a consequence of recording students on the list of students filled out for every class, each 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 several teachers assigned, 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 below.

Table 6: Overview of respondents, instruments, modes and contents of the first wave (2011) of starting cohort 3.

| Respondent | Mode (Instrument) | NEPS Unit | Contents |
|--|--|--|--|
| <i>Target Person</i> (5 th grader) | Paper-based Test | pillar 1 | domain-general cognitive function, reading-p ¹⁾ , reading speed, reading, mathematics-p ¹⁾ |
| | | stage 4 | orthography |
| | PAPI (32,44,45,46) | students' questionnaire | |
| | | pillar 2 | HLE ²⁾ (homework), informal LE (sport lessons); transition, ICT |
| | | pillar 3 | sociodemographics, cultural capital, educational aspirations |
| | | pillar 4 | sociodemographics, language use |
| | | pillar 5 | satisfaction, health |
| | | AG ISM ³⁾ | academic self-concept, self-esteem |
| <i>Parent of child</i> | CATI (31) | stage 4 | reading engagement, quality of instruction, questions about orthography |
| | | pillar 2 | nonformal LE (private lessons) |
| | | pillar 3 | school history of target person, cultural capital |
| | | pillar 4 | language use |
| | | pillar 5 | health |
| <i>Teacher</i> | PAPI (34) | stage 4 | German lessons, parental support |
| | | class teacher questionnaire: questions to the class teacher/class | |
| | | pillar 2 | structural characteristics, composition, equipment |
| | | pillar 3 | social composition |
| | PAPI (35) | pillar 4 | availability and usage of German language and L1 support programs |
| | | stage 4 | quality of instruction, questions about orthographic instruction |
| | PAPI (33) | general teacher questionnaire | |
| | | pillar 2 | participation in further education, pedagogic orientations |
| | | pillar 3 | sociodemographics |
| | | pillar 4 | migration background, L1 |
| <i>Principal</i> | PAPI (36) | pillar 5 | occupational choice, formal qualifications |
| | | stage 4 | workload, cooperation, satisfaction |
| | | pillar 2 | structural characteristics, composition, equipment, staff, quality management |
| | | pillar 3 | sociodemographics, social composition |
| | | pillar 4 | migration background of head, ethnic composition, support programs for migrants |
| | list of students (paper based) list of teachers (paper based) | pillar 5 | composition |
| | | stage 6 | support programs for transition to job |
| | | "Schülerliste": Basic sociodemographics to children (participants & non-participants), assignment to class "Lehrerliste": Assignment of teachers to classes | |

1) -p means procedural metacognition. 2) "H" is used for home; „LE“ is used for learning environment. 3) NEPS workgroup Interest, Self-Concept, Motivation.

4 Data Structure

4.1 Overview

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

All cross-sectional files are stored in wide-format. That means one record represents one respondent at the wave denoted with *x* in the filename. For instance, the file *xEducator* records data from all educators responding to the educator's questionnaire. For episode data, usually collected retrospectively using iterative sets of questions, we provide so-called spell files that are prefixed by a "sp". An example in starting cohort 3 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 7 provides an overview to the data files of the first wave scientific use file of Starting Cohort 3.

Note, that since school cohorts of the NEPS have a multi-level and multi-informant design there are multiple identifiers of persons and entities 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 child's parent. 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 competence test group (ID_tg in *CohortProfile*) and for marking an interviewer in the parent interviewer (ID_int in *ParentMethods*). However, these IDs are not relevant for data merging and for most empirical applications negligible.

Table 7 Summary of data files in 1st wave of starting cohort 3 (5th Graders)

| File | Unique Identifier | Content | N | Availability* |
|---|-------------------|--|-------|---------------|
| Para data and linkage | | | | |
| CohortProfile | ID_t | Para data on all children of the sample. | 6112 | D,R,O |
| ParentMethods | ID_p | Para data from the parent CATI. Includes all parents that could be contacted in the field. | 4753 | D,R,O |
| Survey and test data collected at school | | | | |
| xTarget | ID_t | Data from students | 5753 | D,R,O |
| xTargetCompetencies | ID_t | Competence test data on children | 5208 | D,R,O |
| xEducator | ID_e | Data from Educators | 680 | D,R,O |
| xiInstitution | ID_i | Data from the principals of Kindergartens | 205 | R,O |
| xCourseClass | ID_cc ID_e | Para data and data from the Educators on their classes | 624 | D,R,O |
| xCourseGerman | ID_cg ID_e | Para data and data from the Educators on their German classes | 569 | D,R,O |
| xCourseMath | ID_cm ID_e | Para data from the Educators on their maths classes | 551 | D,R,O |
| Data from parents | | | | |
| xParent | ID_p | Most of data from parent's CATI | 4151 | D,R,O |
| spSchool | ID_p spell | Student's schooling spells. School history collected in the parent interview, but stored in spell format | 8706 | D,R,O |
| spGap | ID_p spell | Data on gaps in event time history provided by parent's CATI | 126 | D,R,O |
| Regional data | | | | |
| xiInstitutionRegioInfas | ID_i regio | Regional data to enhance analyses taking into account the institution | 1164 | O |
| xTargetRegioInfas | ID_t regio | Regional data on the student's place of residence | 19060 | O |

* Availability of file: D = download, R = RemoteNEPS, O = on-site.

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 SC3 these are all students, whose parents agreed to their child's participation. The file has a long format structure with variables *ID_t* and *wave* identifying a row uniquely. For each wave and student it contains information on sample assignment, data availability, and participation status. Furthermore, data from the list of students ("Schülerliste", filled out by the school coordinators, see sections 3.3.5 and 11.2) are stored in this file. This is valuable data since due to nonresponse only for a subset of students data from parental interviews is available.

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. Finally, weighting variables are stored in file *CohortProfile*. In general, we strongly recommend using this file as a starting point of any analysis (see syntax examples in section 0).

In Figure 2 you can see a data example of the *CohortProfile* file. There are five students. According to the records from the list of students they are born between 1998 and 2000 (variable *tx8050y*), three students are male and two are female (variable *tx80501*). For two students there are no assignment to parents (code -55 in *ID_p*), because parents gave no consent or were not able to be contacted in the field work. The variable *stratum* identifies the sample stratum (here school form) to which the student belongs.

Figure 2 Data example of data file *CohortProfile*

| ID_t | wave | stratum | ID_p | ID_cc | ID_cg | ID_cm | tx80220 | tx80524 | tx80501 | tx8050y |
|---------|------|---------|---------|------------|------------|------------|---------|---------|---------|---------|
| 4005733 | 1 | 4 | -55 | 1000364103 | 1000364103 | 1000364103 | 1 | 1 | 1 | 1999 |
| 4005734 | 1 | 5 | -55 | 1000414104 | 1000414104 | 1000414104 | 1 | 1 | 1 | 1998 |
| 4005735 | 1 | 6 | 1001879 | 1000352103 | 1000352103 | 1000352103 | 1 | 1 | 2 | 2000 |
| 4005736 | 1 | 2 | 1001814 | 1000333102 | 1000333102 | 1000333102 | 1 | 1 | 2 | 1999 |
| 4005737 | 1 | 5 | 1001342 | 1000414104 | 1000414104 | 1000414104 | 1 | 1 | 1 | 1998 |

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 5th grade teaching in German and mathematics is usually organized in a class rather than a course context. This means that the teacher who provides the information about the class will also provide the information about the teaching in German. However, there is a small amount of cases in which a course structure applies. Therefore, in order to have a generalized and flexible data format we organized the data in a course format. Every “course” (class, German, or mathematics) 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 mathematics (*xCourseMath*; i.e. a questionnaire about mathematics was not administered in the first wave). 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.

Furthermore, it links a course to its (potentially multiple) teacher(s) in a simple long format with one teacher and one course per wave. 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”). Note that this multiple course-teacher assignment in technical terms represents an n-to-m relationship between courses and teachers.

To simplify this n:m-relationship we provide a generated dummy variable *ex20100* indicating a default course-teacher assignment. This default assignment is derived according to an abstract data edition rule: The teacher validly answering most questions about the course is determined as default, and flagged with the value 1. If such a person cannot be determined, a default teacher is drawn randomly. It should be appropriate to use this default educator for most analyses (for an illustrative application see example in section 7.4 on page 21f). Be aware that this is just a technical definition—in the context of your analyses it might be reasonable to rely on more substantial selection criteria.

Figure 3 shows an exemplary snapshot from the *xCourseClass* data file. Depicted are six records pertaining to four classes (more generally “courses”, see above) identified by *ID_cc*. The educator identifier variable *ID_e* references those teachers, who were assigned as class teachers to the respective class. Each row represents a unique teacher-class relation. Variable *ex20101* indicates whether the teacher participated at the class teacher questionnaire. As noted above, *ex20100* indicated a “default” teacher assignment as defined by the NEPS data edition. Eventually, *e451010* and *e229400_D* are two variables from the class teacher questionnaires just selected arbitrarily for illustrative reasons. 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. Since, this teacher is the only teacher assigned to this class, she is trivially set to be the “default” class-teacher assignment. For the second class, two teachers were assigned and both participated in the class teacher questionnaire. However, the second one (*ID_e*=1001538) was elected by data edition to be the “default” teacher, because she provides more non-missing information on the class (note that though *ID_e*=1001815 participated at the questionnaire there is “-90” item-response in the selected questionnaire variables). With regard to the third class, there were also two teachers assigned, however, none of them filled out a class teacher questionnaire. In absence of further information, the first teacher was set as default by random. The forth class is an “ideal” case. One teacher was assigned, who participated and provided non-missing information.

The same logic applies for *xCourseGerman* and *xCourseMath*, however, the latter does not include questionnaire variables, since there was no math teacher questionnaire in the first wave.

Figure 3 Data example of data file *xCourseClass*

| ID_e | ID_cc | ID_i | ex20100 | ex20101 | e451010 | e229400_D |
|---------|------------|---------|---------|---------|---------|-----------|
| 1005823 | 1000157101 | 1000157 | 1 | 0 | -56 | -56 |
| 1001851 | 1000177101 | 1000177 | 0 | 1 | -90 | -90 |
| 1001538 | 1000177101 | 1000177 | 1 | 1 | 8 | 3 |
| 1005702 | 1000202101 | 1000202 | 1 | 0 | -56 | -56 |
| 1006260 | 1000202101 | 1000202 | 0 | 0 | -56 | -56 |
| 1001525 | 1000205103 | 1000205 | 1 | 1 | 2 | 5 |

4.2.3 *xTarget*: student's answers

In general, students are the target persons in the school cohorts of the NEPS. Within the context of the 5th grade, questionnaire data on students was collected from them via a PAPI questionnaire. The coded data is stored in the file *xTarget* following a simple cross-sectional structure (i.e. 1 row = 1 student). Students are identified by *ID_t*. Note that *xTarget* only includes responding students of the panel sample, whilst *CohortProfile* includes all students of the panel sample.

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 they 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.

4.2.5 *xEducator*: data on teachers

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*: answers from principals

Answers to the principal questionnaire provided by the school principals are stored in *xInstitution*. Analogously to *xEducator* and *xTarget* only responding principals are included. Note, due to data protection issues, this file is not available in the download version of the dataset.

4.2.7 Datafiles 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 to the interview). Moreover, questionnaire data consists of cross-sectional data (at time of interview) as well as of retrospective data requiring a spell data format for convenient use. While one

integrated CATI instrument has been employed for starting cohort 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 parental interviews in several data files. Table 16 in the appendix (p. 37) shows which files with parent data are available by starting cohort. In the following, we describe parent files available for first wave of starting cohort 3.

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*.

spSchool

Specifically for starting cohort 3, the target student's schooling history is surveyed retrospectively. 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 one school episode reported by a specific parent (see data example in Figure 4). Thus, a row is uniquely identified by *ID_p* and the counter variable *spell*. Each spell has spell times, i.e. a start date and an end date each of with 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*). In addition to the spell times reported originally in the questionnaire module, there are spell times potentially corrected for inconsistencies and wrong temporal ordering in the check module (see section 3.3.3). The 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 which times to use is up to you as a researcher.

Figure 4 Data example of data file *spSchool*

| <i>ID_p</i> | <i>spell</i> | <i>p723080</i> | <i>p72301m</i> | <i>p72301y</i> | <i>p72302m</i> | <i>p72302y</i> | <i>p723110</i> |
|-------------|--------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1000804 | 1 | 1 | 9 | 2006 | 7 | 2010 | . |
| 1000804 | 2 | 5 | 9 | 2010 | 2 | 2011 | 1 |
| 1000805 | 1 | 1 | 9 | 2005 | 2 | 2011 | 1 |
| 1000806 | 1 | 1 | 9 | 2005 | 8 | 2009 | . |
| 1000806 | 2 | 5 | 8 | 2009 | 2 | 2010 | . |
| 1000806 | 3 | 5 | 9 | 2010 | 2 | 2011 | 1 |

spGap

Contains gap episodes that have been reported in the check module while checking the consistency of the school history. For example, gaps were filled up if the student was ill or away on holidays. Since both contain spells in a long format, the data structure of file *spGap* is analog to the structure of file *spSchool*.

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. However, only for a subset of those cases an interview was effectively realized. Thus, *ParentMethods* does include more cases than *xParent* does. A significant amount of the surplus parents in *ParentMethods* are just temporary dropouts (i.e. will be 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.⁴ You can use the categorical variable *px80220* for analyzing the participation status of parents. The figure below gives an exemplary snapshot from the *ParentMethods* file. It shows some para data on four parents, of whom three gave interviews while one could not be interviewed and, thus, is coded as temporary dropout.

Figure 5 Data example of data file *ParentMethods*

| ID_p | ID_t | wave | intm | inty | px80220 |
|---------|---------|------|------|------|---------|
| 1000804 | 4007319 | 1 | 2 | 2011 | 1 |
| 1000979 | 4006304 | 1 | 2 | 2011 | 1 |
| 1006589 | 4008048 | 1 | 5 | 2011 | 1 |
| 1016609 | 4007059 | 1 | . | . | 2 |

For convenience, the variable *ID_t* is included for mapping parents to targets, i.e. students. Note, that the mapping of children to parents— i.e. *ID_t* to *ID_p*—is already established in *CohortProfile*. Due to technical reasons, each *ID_t* is assigned an own *ID_p* (i.e. there is a one-to-one relationship). Consequently, since they have different values for *ID_t* siblings in the sample have different *ID_p* even if they live in the same parental context.

4.2.8 Regional data: *xInstitutionRegioInfas* and *xTargetRegioInfas*

Fine-grained regional data is provided in the data files *xTargetRegioInfas* and *xInstitutionRegioInfas*. These files have been generated from the *infas geodaten* data-

⁴ Nevertheless they are included in the file since they gave their consent prior at the time of invitation.

base.⁵ Both comprise geographical information on four regional levels (coded in variable *regio*): municipality, postal code, quarters (living areas), and street sections. These data were linked to each target and institution by geocoding 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 available in Supplement “RegioInfas / infas geodaten” (see 1.1).

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 on an ex-post-factum basis which given information is valid and which invalid (the filter question was wrongly answered versus the questions to skip were wrongly answered). Therefore, the NEPS data center 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 PAPI data.

The cleaning syntax is part of Supplement “Syntax package” and can be downloaded from the SC3 data section on our web portal (see section 0). If you are using RemoteNEPS or the onsite-version of the SUF, you additionally will find the syntax files in the directory “Z:\Public” as soon as they are finalized. The syntax files’ names 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 corresponding directory. It describes the cleaning procedure in detail. Also see example 5 in section 7.5 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 8 presents an overview of these coded variables. Furthermore, the parental educational information was coded by using the CASMIN and the ISCED-97 classification

⁵ This database is provided by the infas geodaten GmbH, see: <http://www.infas-geodaten.de>

and a metric variable offering the standardized years of occupation. Tables how EGP (cf. Erikson et al. 1979), the BLK (classification of occupations according to Blossfeld, cf. Blossfeld 1985; Schimpl-Neimanns 2003) the ISCED-97 (UNESCO 2006) and CASMIN (Lüttinger & König 1988) classes are coded are presented in the following.

Table 8 Overview of coded variables

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

Table 9 Coding of EGP

| Codes | 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 nicht-manuellen Klasse mit beschränkten Entscheidungsbefugnissen |
| 4 | [IIIb] | Lower Sales-Service | Angestellte der ausführenden nicht-manuellen Klasse mit gering qualifizierten Routinetätigkeiten |
| 5 | [IVa] | Self-employed with employees | Selbständige mit unterstellten Mitarbeitern |
| 6 | [IVb] | Self-employed no employees | Selbständige ohne unterstellte Mitarbeiter |
| 7 | [IVc] | Self-employed 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 10 Coding of BLK

| Codes | 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] | Semi-professions | Semiprofessionen |
| 9 | [PROF] | Professions | Professionen |
| 10 | [EVB] | Common commercial and administrative occupations | Einfache kaufmännische und Verwaltungsberufe |
| 11 | [QVB] | Skilled commercial and administrative occupations | Qualifizierte kaufmännische und Verwaltungsberufe |
| 12 | [MAN] | Manager | Manager |

Table 11 Coding of ISCED-97

| Codes | ISCED-97 | | |
|-------|----------|--|---|
| | Key | English | German |
| 0 | 0A/1A | Inadequately 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 specialized 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 12 Coding of CASMIN

| Codes | CASMIN | | |
|--------------|---------------|---|--|
| | Key | English | German |
| 0 | 1a | Inadequately 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 *CohortProfile* dataset. Given the quite complex structure of the sample of the 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. Adjusted design weights provided are named *weight_design*. Note that also standardized weights with mean one are provided, which are often used in statistical analysis. These are named *weight_design_std*. Information on weight construction and how to use them can be found in the technical report on weighting (Supplement “Weighting”) and the examples section (see section 7.6).

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.⁶ In future releases of this manual we will extend the examples by code in R and SPSS.

⁶ In our Stata examples we make use of the user-written command “fre”, a powerful alternative to “tabulate” for displaying one-way frequency tables. It can be easily installed by typing “ssc install fre” into the Stata console. Acknowledgment goes to Ben Jann (2007), who developed this 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.

Example 1 in Stata

```

/*****
Example 1: Using CohortProfile

Procedure:
- drop cases of sample of special needs students
- check data availability of the sample of regular students

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/

* load CohortProfile
use "SC3_CohortProfile_${version}", clear

* for all variables recode NEPS missings to Stata missings
nepsmiss _all

* drop sample of special needs students
drop if stratum==7

* count cases of sample of regular students
count

* Check status of survey participation
*      (participation, here defined as
*      having participated in tests or
*      having target data)
fre tx80220

* Exploit data from the students' list included in CohortProfile
* sex of child
fre tx80501
* year of birth of child
fre tx8050y

* check data availability: test data and parent data
tabulate tx80522 tx80523, cell miss

```

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.

Example 2 in 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. analyze combined data

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/

* load CohortProfile as master dataset
use "SC3_CohortProfile_${version}", clear

* merge file xTarget
* (merge row by row uniquely identified by ID_t,
* keep matched and un-matched cases of CohortProfile)
merge 1:1 ID_t using "SC3_xTarget_${version}", nogenerate keep(master matched)

* merge file xParent (consider missing parent IDs for some children)
nepsmis ID_p //recode NEPS missings to Stata missings
* since there are missings in ID_p there is a n:1 relationship for merge
merge n:1 ID_p using "SC3_xParent_${version}", nogenerate keep(master matched)

* recode missings
nepsmis _all

* check agreement on student's sex between
* data from list of students (tx80501 in CohortProfile,
* available for nearly all students) and data from the student's
* questionnaire (t700031 in xTarget, available only for responding
* students, of course)
tabulate tx80501 t700031, cell // ~99% agreement
* agreement between students' and parents' responses on sex
tabulate t700031 p700010, cell // ~99% agreement

```

7.3 Example 3 – Indirectly matching different datasets

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 some variation on the level of classes. For simplicity, say class-level information can be measured by characteristics of the very teachers who take care of the respective class.

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 also reflects the n:m relationship between teachers and courses: One course can consist of multiple teachers and one teacher can be related to multiple courses. By merging the teacher's information contained in *xEducator* to *xCourseClass* and afterwards merging this newly generated information to the information generated in the first step, you can easily draw upon the rich multilevel information available in the data of starting cohort 3. Everything you have to do after this procedure is to aggregate the information provided by the teachers according to a rule. In this example we generate a class level variable “mean age of class teacher(s)” by averaging over the estimated age of all class teachers that are assigned to a class. Subsequently, this generated variable is merged to students in *CohortProfile*.

Example 3 in Stata

```

/*****
* Example 3: Indirectly match teachers to students
              by preparing course-level data
Procedure
1. prepare course-level data: "(mean) age of class teacher(s)"
2. load CohortProfile
3. merge prepared class-level data
4. analyze data

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/

* 1) Prepare course-level data: "(mean) age of class teacher(s)"
* load xCourseClass file
use "SC3_xCourseClass_${version}", clear

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

* merge age of assigned teachers from xEducator
merge n:1 ID_e using "SC3_xEducator_${version}", ///
              nogenerate keep(master matched) keepusing(e76212y_D)
nepsmiss e76212y_D

* replace categories with class center
recode e76212y_D (1=1945) (2=1955) (3=1965) (4=1975) (5=1985)

* estimate age of class teachers (roughly!)
generate eduage = 2010 - e76212y_D

* keep only relevant variables
keep ID_cc ID_e eduage

* re-organize dataset: reshape to wide over teachers
bysort ID_cc (ID_e): generate number = _n
keep ID_cc number eduage
reshape wide eduage, i(ID_cc) j(number)

* generate class variable "mage" (mean age of class teacher(s))
egen mage = rowmean(eduage*)
drop eduage*

* temporarily save prepared class file
tempfile teachercontext
save "`teachercontext'", replace

* 2) Load CohortProfile file
use ID_t ID_cc wave stratum using "SC3_CohortProfile_${version}", clear

* 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 "`teachercontext'", nogenerate keep(master matched)

* 4) Analyze data (e.g. inspect distribution of classes' teacher age over children)
summarize mage, detail
* class teacher's mean age by school forms (plus standard deviation)
tabstat mage, by(stratum) stats(mean sd) format(%9.2f)

```

7.4 Example 4 – Prepare multilevel data

In the following example, we will choose a multilevel approach for analyzing reading competencies of students in schools. Note that multilevel modeling is an advanced topic in social science research and real applications should be by far more elaborated than it could be done here. However, the example is intended to illustrate how data from different levels—students, classes, and school level—and from multiple sources—*xTarget*, *xParent*, *xCourseClass*, *xTargetCompetencies*—can be matched together for serious analyses.

Example 4 in Stata

```

/*****
* Example 4: Prepare and analyze multilevel data

Goal:
Regress reading speed on traits of students,
traits of class, and school form.
Account for nested data structure.

1) prepare group level data (dummy for 50% or more migrants in class)
2) prepare data from parental context (highest ISEI)
3) prepare competence data (sum scores of reading speed)
4) prepare students' data
5) integrate data using CohortProfile
6) analyze data (linear mixed model)

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/

* 1) prepare class-level data (dummy for 50% or more migrants in class)
* (resolve n:m relationship between classes and teachers)
* load xCourseClass file
use "SC3_xCourseClass_${version}", clear
* keep "default" teacher assignments (= providing most nonmissing information)
keep if ex20100==1
* keep relevant variables and recode missings
keep ID_cc ID_e e227400 e227401 e451000 e451010
nepsmis _all
* number of students in class
generate nofstud = e227400 + e227401
* calculate fraction of migrants in class
generate migfrac = e451000/nofstud
* generate a dummy for 50% or more migrants in class
generate mig50plus = migfrac>=.5 if !missing(migfrac)
* additionally exploit information from the fuzzy item
replace mig50plus = 1 if inlist(e451010, 4,5,6) & missing(mig50plus)
replace mig50plus = 0 if inlist(e451010, 1,2,3) & missing(mig50plus)
* keep relevant variables
keep ID_cc ID_e mig50plus
* ID_cc now unique
isid ID_cc
* temporarily save
tempfile classcontext
save "`classcontext'", replace

* 2) prepare data from parents (highest ISEI of parental context)
use "SC3_xParent_${version}", clear
nepsmis p731904_g14 p731954_g14
* generate variable for highest ISEI
generate phisei08 = max(p731904_g14,p731954_g14)
keep ID_p phisei08

```

```

* temporarily save
tempfile parentcontext
save "`parentcontext'", replace

* 3) prepare competence data (simplification: sum scores of reading speed)
use "SC3_xTargetCompetencies_${version}", clear
* standardize sum score of vocabulary test
nepsmis rsci_sc3
egen zrs = std( rsci_sc3)
keep ID_t zrs
tempfile zrs
save "`zrs'", replace

* 4) prepare students data
* (scales for pro-social and problematic behavior)
use "SC3_xTarget_${version}", clear
keep ID_t t400000_g1D t700031 t70004y
nepsmis _all
recode t700031 (1=1) (2=0), gen(boy)
recode t400000_g1D (1=1) (0= 0), gen(abroad)
generate age = 2010 - t70004y
keep ID_t boy abroad age
tempfile student
save "`student'", replace

* 5) integrate data using CohortProfile
use "SC3_CohortProfile_${version}", clear
* drop cases from the special needs sample
drop if stratum == 7
nepsmis _all
merge n:1 ID_cc using "`classcontext'", keep(master matched) nogenerate
merge n:1 ID_p using "`parentcontext'", keep(master matched) nogenerate
merge 1:1 ID_t using "`student'", keep(master matched) nogenerate
merge 1:1 ID_t using "`zrs'", keep(master matched) nogenerate

keep ID_t ID_cc ID_i stratum mig50plus phisei08 age abroad boy zrs

* 6) analyze data (linear mixed model)
* do a complete case analysis, listwise deletion of missings
egen nofmiss = rowmiss(ID_t ID_cc ID_i mig50plus phisei08 age abroad boy zrs)
drop if nofmiss > 0
* estimate a linear mixed model that explains variation in zvr by
* student- and class-level variables.
* define a random intercept model with 2 random effects,
* one at the level of institution and one at the level of classes
xtmixed zrs boy age abroad phisei08 mig50plus || ID_i: || ID_cc:
* additionally, control for form of school
xtmixed zrs boy age abroad phisei08 mig50plus i.stratum || ID_i: || ID_cc:
* calculate intraclass correlations (ssc install xtmrho)
xtmrho

```

7.5 Example 5 – Using cleaning syntax for PAPI filtering

In the example we do filter cleaning for the data from teachers. See section 4.3 for details on using the filter question.

Example 5 in Stata

```
/******  
* Example 5: Using filtering syntax  
  
Note: replace ${version} to your file version, e.g. 'R_1-0-0'  
*****/  
  
* load xEducator  
use "SC3_xEducator_${version}", clear  
  
* e.g. variable e537042  
* Before: missing codes -95 (implausible) and  
*         -90 (unknown missing)  
fre e537042  
  
* apply filter syntax  
* (assuming do-file stored in the same directory like the data file)  
quietly: do "SC3_xEducator_${version}.do"  
  
* e.g. variable e537042  
* After: missing codes -95 (implausible),  
*        -90 (unknown missing), and  
*        -99 (filtered)  
fre e537042  
  
* recode missings  
nepsmiss _all
```

7.6 Example 6 – Using weights

This example intends to illustrate how weights provided in CohortProfile can be used. After merging variable *t66000a_g1* (Self concept: German) to the *CohortProfile*-dataset, 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.

Example 6 in Stata

```

/*****
* Example 6: Using weights

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/

* load CohortProfile
use "SC3_CohortProfile_${version}", clear

* merge variable self concept: German from xTarget
merge 1:1 ID_t using "SC3_xTarget_${version}", assert(match master) keep(match)
keepusing(t66000a_g1) nolog generate

* encode missings
nepsmis t66000a_g1 tx80501

* calculate means without weighting
mean t66000a_g1, over(tx80501)

* calculate means using probability weights
mean t66000a_g1 [pweight=weight_design_std], over(tx80501)

* calculate means using sampling weights (svy-command)
svyset ID_i [pweight=weight_design_std], strata(stratum)
svy: mean t66000a_g1, over(tx80501)

```

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 re-identification of persons!
- immediately inform NEPS of any accidental re-identification 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 user service (see section 0). Also, if you encounter any security leaks regarding data protection and data security, or any data quality deficiencies please inform the NEPS user service (see section 0).

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 cannot 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!

- if you encounter problems or even errors in the data please contact the user service of the NEPS!

9 Tools for Stata users

Our Stata files offer variable labels and value labels both in German and in English. You can easily switch between these languages using the `label language` command.

```
label language en
label language de
```

In Addition, NEPS datasets are signed using Stata's `datasignature`. You can check if you are working on unchanged NEPS SUF datasets using the command `datasignature confirm` at any time.

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

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

We try to fix any reported bugs in the `nepstools` frequently. Thus, you should – once installed – make sure that you have the most recent version by executing following command:

```
adoupdate nepstools, update
```

Please read the help files delivered with `nepstools` for a detailed documentation of usage:

```
help nepstools
help nepsmis
help infoquery
```

nepsmis: Recoding missing values

This program automatically recodes and labels all missing values into extended missing values (.a, .b, etc.). In this example, we run `nepsmis` on the variable `t731454`, decoding all negative values (-54, -97, -98) into Stata's extended missings (.c, .b, .a).

```
nepsmis t731454
```

| ID_t | wave | t731454 |
|---------|------|---------|
| 8010851 | 2 | -97 |
| 8012254 | 1 | -54 |
| 8002388 | 2 | -98 |
| 8012254 | 2 | 5 |
| 8002388 | 1 | 1 |

| ID_t | wave | t731454 |
|---------|------|---------|
| 8010851 | 2 | .b |
| 8012254 | 1 | .c |
| 8002388 | 2 | .a |
| 8012254 | 2 | 5 |
| 8002388 | 1 | 1 |

We generally recommend running `nepsmis` on all variables (`nepsmis _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 t405060
```

```
query result for variable t405060:
```

```
t405060[questiontext_de]:
```

```
Wo ist Ihre Mutter (Stiefmutter / diese Person) geboren?
```

```
t405060[questiontext_en]:
```

```
Where was your mother (stepmother/ this person) born?
```

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.

<https://www.neps-data.de/de-de/datenzentrum/forschungsdaten/startkohorteklasse5>
(in German)

<https://www.neps-data.de/en-us/datacenter/researchdata/startingcohort5thgrade>
(in English)

For further support, please contact the NEPS data center:

Web:

<https://www.neps-data.de/de-de/datenzentrum/kontaktzentrum>
(in German)

<https://www.neps-data.de/en-us/datacenter/contactdatacenter>
(in English)

E- Mail:

userservice.neps@uni-bamberg.de

Phone:

+49 951 8633511 (Mo-Fr 10:00-12:00 and 14:00-16:00)

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.

11 Appendix

11.1 Characteristics of the survey of students with special educational needs

11.1.1 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.

11.1.2 Starting cohort 3

In wave one of the five graders cohort, the questionnaire for students of the SEN sample was shorter than the questionnaire of the regular students' sample. Moreover, contrary to the regular students special needs students had two testing days.

The questionnaire was split up in a first part asking of basic socio-demographic information (labeled as "BI" in the following) and in a second part containing four modules with different topics (M1 to M4). The first part was administered at the end of first testing day. The second part was administered in two split versions with reversed sequence at the end of the second testing day. As a results of this design, for students who were absent on the first testing day (e.g. due to illness) but not on the second, there is only information available to M1-M4 but not to BI (and, of course, vice versa). The design is summarized in the following table; in the *xTarget* dataset, it is represented by the variables *tx80211_w1* and *tx80211_w2*. Details of question ordering is given in Table 13.

Table 13 Overview to questionnaire rotations for special needs students on 1st wave of SC3

| | Testing day 1 | Testing day 2 |
|---------|----------------------------------|---|
| Split 1 | BI (Booklet 1; Instrument 44) | M1 + M2 + M3 + M4 (Booklet 2; Instrument 45) |
| Split 2 | BI (Booklet 1) | M4 + M3 + M2 + M1 (Booklet 3; Instrument 46) |

Table 14 Sequence of questions in different rotations in the questionnaires of SEN students in 1st wave of SC3

| | Instruments of survey of special needs students (A56) | | | | |
|---|---|-----------------------|-----------|-----------|--------|
| Number of question in A28 (reg. students) | Question included in A56 | Number of question in | | | Module |
| | | Booklet 1 | Booklet 2 | Booklet 3 | |
| 1 | x | 1 | | | BI |
| 2 | x | 2 | | | BI |
| 3 | x | 3 | | | BI |
| 4 | x | 4 | | | BI |
| 5 | x | 5 | | | BI |
| 6 | | | | | |
| 7 | | | | | |
| 8 | | | | | |
| 9 | x | 6 | | | BI |
| 10 | x | 7 | | | BI |
| 11 | x | 8 | | | BI |
| 12 | x | | 8 | 12 | M2 |
| 13 | x | 9 | | | BI |
| 14 | x | 10 | | | BI |
| 15 | x | 11 | | | BI |
| 16 | x | 12 | | | BI |
| 17 | x | 13 | | | BI |
| 18 | x | 14 | | | BI |
| 19 | x | 15 | | | BI |
| 20 | | | | | |
| 21 | | | | | |
| 22 | | | | | |
| 23 | | | 7 | 11 | M2 |
| 24 | x | | | | |
| 25 | x | 16 | | | BI |
| 26 | x | 17 | | | BI |
| 27 | | | 6 | 10 | M2 |
| 28 | x | 18 | | | BI |
| 29 | | | | | |
| 30 | | | | | |
| 31 | | | | | |
| 32 | x | | 9 | 6 | M3 |
| 33 | x | | 10 | 7 | M3 |
| 34 | x | | | | |
| 35 | x | | 11 | 8 | M3 |
| 36 | x | | 12 | 9 | M3 |
| 37 | x | | 1 | 13 | M1 |
| 38 | | | 2 | 14 | M1 |
| 39 | | | | | |
| 40 | | | 3 | 15 | M1 |
| 41 | | | | | |
| 42 | | | 4 | 16 | M1 |
| 43 | | | 5 | 17 | M1 |
| 44 | | | | | |
| 45 | x | | 13 | 1 | M4 |
| 46 | x | | 14 | 2 | M4 |
| 47 | x | | 15 | 3 | M4 |
| 48 | x | | 16 | 4 | M4 |
| 49 | | | 17 | 5 | M4 |
| 50 | | | | | |
| 51 | | | | | |
| 52 | | 19 | | | BI |
| 53 | x | 20 | | | BI |
| 54 | x | 21 | | | BI |
| 55 | x | 22 | | | BI |
| 56 | x | 23 | | | BI |
| 57 | x | 24 | | | BI |
| 58 | x | 25 | | | BI |

11.2 “Schülerliste” – List of students

The list of students (see Figure 7) was divided into a left, 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 anonymised 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* (see Figure 6). 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.

Figure 6 Right section of the list of students ("Schülerliste") and variables in CohortProfile

✂

Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 5

Schulname: Musterschule

Schul-ID: 255201

Klasse: 5c

Testgruppen-ID: 25520131

Legende:
Welche Empfehlung für die weiterführende Schule hat der Schüler durch die Grundschule erhalten?

HS = Hauptschule; RS = Realschule; GY = Gymnasium; SMB = Schule mit mehreren Bildungsgängen (z. B. kombinierte Haupt- und Realschule, Oberschule, Mittelschule, Regelschule, Regionalschule, Realschule plus, Sekundar- oder erweiterte Realschule); IGS = Integrierte Gesamtschule; FOE = Förderschule; SON = Sonstige.
Wenn keine Empfehlung ausgestellt bzw. kein Übertrittszeugnis beantragt wurde (betrifft bspw. Bayern) oder Sie keine Einschätzung vornehmen können, bitte NB (für nicht beantragt) eintragen.

| [13] | [14] | [15] | [16] | | [17] | [18] | | [19] | | | | [20] |
|-------------------|--|--|--------------------------------|--------------------------------------|---|----------------|------------------|-------------------|-----------------|----|----|-------------|
| Elterngenehmigung | Obertrag aus Spalte [10] Geschlecht | Obertrag aus Spalte [11] Geburtsmonat und -jahr | Einschätzung der Lesekompetenz | Einschätzung der Mathematikkompetenz | Obertrag aus Spalte [7] Übergangsempfehlung der Grundschule (s. Legende oben) | Instrument-Nr. | | Teilnahmezustatus | | | | Schüler-ID |
| (ja/nein) | (m/w) | (MM.JJ) | (Skala 1 - 6) | (Skala 1 - 6) | FB | TH | Haupttest (TAVE) | | Nachtest (TAVE) | | | |
| | | | | | | | | TH | FB | TH | FB | |
| ja | m | 12.94 | 3 | 2 | GY | 1 | 1 | | | | | 2552019-001 |
| ja | w | 03.95 | 2 | 2 | GY | 2 | 2 | | | | | 2552019-002 |
| nein | | | | | | 1 | 3 | | | | | 2552019-003 |
| nein | | | | | | 2 | 4 | | | | | 2552019-004 |
| ja | | | | | | 1 | 1 | | | | | 2552019-005 |
| nein | | | | | | 2 | 2 | | | | | 2552019-006 |
| ja | m | 01.95 | 4 | 5 | GY | 1 | 3 | | | | | 2552019-007 |
| ja | w | 02.95 | 3 | 1 | GY | 2 | 4 | | | | | 2552019-008 |
| nein | | | | | | 1 | 1 | | | | | 2552019-009 |
| ja | w | 10.94 | 3 | 3 | GY | 2 | 2 | | | | | 2552019-010 |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |

↓

tx80501
(sex)

↓

tx8050m/y
(date of birth)

→ Dataset CohortProfile

Figure 7 "Schülerliste" of grade 5

**Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 5**

Schulname: Musterschule

Schul-ID: 255201

Klasse: 5c

Testgruppen-ID: 25520131

**Haupterhebung K5/K9
Angaben zur Grundgesamtheit**

Information im Klartext: Bundesland, Schulform, Jahrgang

**Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 5**

Schulname: Musterschule

Schul-ID: 255201

Klasse: 5c

Testgruppen-ID: 25520131

Legende:

Welche Empfehlung für die weiterführende Schule hat der Schüler durch die Grundschule erhalten?

HS = Hauptschule; RS = Realschule; GY = Gymnasium; SMB = Schule mit mehreren Bildungsgängen (z. B. kombinierte Haupt- und Realschule, Oberschule, Mittelschule, Regelschule, Regionalschule, Realschule plus, Sekundar- oder erweiterte Realschule); IGs = integrierte Gesamtschule; FÖE = Förderschule; SON = Sonstige.
Wenn keine Empfehlung ausgestellt bzw. kein Übertrittszeugnis beantragt wurde (beintr. bspw. Bayern) oder Sie keine Einschätzung vornehmen können, bitte NB (für nicht beantragt) eintragen.

| [1] | [2] | [3] | [4] | [5] | [6] | | [7] | [8] | [9] | [10] | [11] | [12] | [13] | [14] | [15] | [16] | | [17] | | [18] | | [19] | | | [20] |
|----------|--|-------------|---|---|---|---|---|--------------------------------------|---|---------------------|-----------------------------------|------------------------------------|--------------------------------|---------------------|-----------------------------------|---|---|--|----------------|------|------------------|------|-----------------|----|-------------|
| Lfd. Nr. | Name des Schülers (Nachname, Vorname) | Schüler-ID | Staatsangehörigkeit 1= deutsch 2= nicht-deutsch | Deutsch als vorwiegend zu Hause gesprochene Sprache (ja/nein) | Einschätzung der Lesekompetenz (Skala 1 - 6) | Einschätzung der Mathematikkompetenz (Skala 1 - 6) | Übergangsempfehlung der Grundschule (s. Legende oben rechts) | Sonderpäd. Förderbedarf (ja/nein) | Diagnostizierte Leserechtschreibschwäche (ja/nein) | Geschlecht (m/w) | Geburtsmonat und -jahr (MM.JJ) | Teilnahmebereitschaft (ja/nein) | Elterngenehmigung (ja/nein) | Geschlecht (m/w) | Geburtsmonat und -jahr (MM.JJ) | Einschätzung der Lesekompetenz (Skala 1 - 6) | Einschätzung der Mathematikkompetenz (Skala 1 - 6) | Übergangsempfehlung der Grundschule (s. Legende oben) | Instrument-Nr. | | Haupttest (TAVE) | | Nachtest (TAVE) | | Schüler-ID |
| | | | | | | | | | | | | | | | | | | | FB | TH | TH | FB | TH | FB | |
| 001 | Akcan, Deniz | 2552019-001 | 2 | nein | 3 | 2 | GY | nein | nein | m | 12.94 | ja | ja | m | 12.94 | 3 | 2 | GY | 1 | 1 | | | | | 2552019-001 |
| 002 | Böning, Claudia | 2552019-002 | 1 | ja | 2 | 2 | GY | nein | nein | w | 03.95 | ja | ja | w | 03.95 | 2 | 2 | GY | 2 | 2 | | | | | 2552019-002 |
| 003 | Freese, Verena | 2552019-003 | 1 | ja | 5 | 3 | GY | nein | ja | w | 08.95 | nein | nein | | | | | | 1 | 3 | | | | | 2552019-003 |
| 004 | Frisch, Lukas | 2552019-004 | 1 | ja | 4 | 4 | GY | nein | nein | m | 11.94 | nein | nein | | | | | | 2 | 4 | | | | | 2552019-004 |
| 005 | Günther, Marisa | 2552019-005 | 1 | ja | 1 | 2 | GY | nein | nein | w | 12.95 | nein | ja | | | | | | 1 | 1 | | | | | 2552019-005 |
| 006 | Halberstadt, Julius | 2552019-006 | 1 | ja | 2 | 3 | GY | nein | nein | m | 02.95 | nein | nein | | | | | | 2 | 2 | | | | | 2552019-006 |
| 007 | Liebig, Sebastian | 2552019-007 | 1 | ja | 4 | 5 | GY | ja | ja | m | 01.95 | ja | ja | m | 01.95 | 4 | 5 | GY | 1 | 3 | | | | | 2552019-007 |
| 008 | Landmann, Anne-Kathrin | 2552019-008 | 1 | ja | 3 | 1 | GY | nein | nein | w | 02.95 | ja | ja | w | 02.95 | 3 | 1 | GY | 2 | 4 | | | | | 2552019-008 |
| 009 | Marx, Alexander | 2552019-009 | 1 | ja | 3 | 4 | GY | nein | nein | m | 12.94 | nein | nein | | | | | | 1 | 1 | | | | | 2552019-009 |
| 010 | Müller, Renée | 2552019-010 | 2 | ja | 3 | 3 | GY | nein | nein | w | 10.94 | ja | ja | w | 10.94 | 3 | 3 | GY | 2 | 2 | | | | | 2552019-010 |
| | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | |

11.3 “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 lists as illustrated below.

Figure 8 List of teachers for schools where math and German is taught in a class context

| NEPS-Haupterhebung K5/K9 Herbst/Winter 2010 | | | NEPS-Haupterhebung K5/K9 Herbst/Winter 2010 | | | | | | | | | | | |
|---|-------------------------------|-----------------------|--|----|----|-----------|----|----|-----------|----|----|-----------|----|----|
| Schulname: Schul-ID: | | | Schulname: Schul-ID: | | | | | | | | | | | |
| Lehrerliste | | | Lehrerliste | | | | | | | | | | | |
| Bitte listen Sie hier alle Klassen-, Deutsch- und Mathematiklehrkräfte, die in den NEPS-Klassen unterrichten. | | | 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 <u>rechten Abschnitt</u> der ausgefüllten Lehrerliste bis spätestens XX.XX.2010 an das DPC, Fax-Nummer: 040 48500-501. | | | | | | | | | | | |

Figure 9 List of teachers for schools where math and German is taught as courses.

| NEPS-Haupterhebung K5/K9 Herbst/Winter 2010 | | NEPS-Haupterhebung K5/K9 Herbst/Winter 2010 | | | | | | |
|--|-------------------------------|---|-------------------------|----|----|----|--|--|
| Schulname: | | Schulname: | | | | | | |
| Schul-ID: | | Schul-ID: | | | | | | |
| Lehrerliste | | Lehrerliste | | | | | | |
| <i>Bitte listen Sie hier alle Klassen-, Deutsch- und Mathematiklehrkräfte, die in den NEPS-Klassen unterrichten.</i> | | <i>Bitte geben Sie hier an, welche Funktion die in Spalte [2] genannte Person in der jeweiligen Klasse (Spalten [4] bis [6]) innehat.</i> | | | | | | |
| [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 <u>nur</u> die lfd. Nr. aus der Kursliste eintragen) | (bitte <u>nur</u> 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.

11.4 Modules of parental interview

Table 15 Modules of parental interview (CATI, 1st wave) by starting cohort

| # | CATI modules | Starting Cohort | | |
|----|---|-----------------|---|---|
| | | 2 | 3 | 4 |
| 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 16 Datasets containing data from the parent interview by starting cohort (1st wave)

| Data files | Unique Identifiers | Content | Starting Cohort | | |
|---------------|-------------------------|---|-----------------|---|---|
| | | | 2 | 3 | 4 |
| xParent | ID_p | Most data from CATI interview stored in cross-sectional format (1 row = 1 responding parent) | x | x | x |
| spChildCare | ID_p sptype spell | History of child care for target child in spell format (1 row = 1 care spell) | x | | |
| spSchool | ID_p spell | History of schooling for target child in spell format (1 row = 1 school spell) | | x | x |
| spGap | ID_p spell | Gap episodes in school history of target child in spell format (1 row = 1 gap spell) | | x | x |
| 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; | x | x | x |

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

Illustration 1 Exemplary test item for listening in grade 9

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‘.

















| | | | | |
|---|---|---|---|---|
| A |  |  |  |  |
| | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| B |  |  |  |  |
| | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| C |  |  |  |  |
| | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| D |  |  |  |  |
| | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Illustration 2 Exemplary test item for ICT competence

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?

Ergebnisse finden

mit **allen** Wörtern

10 Ergebnisse ▼

mit der **genauen Wortgruppe**

mit **irgendeinem** der Wörter

ohne die Wörter

Sprache Antwortseiten, geschrieben in beliebige Sprache ▼ **①**

Region Gesuchte Seiten befinden sich in: alle Regionen ▼ **②**

Dateiformat Ausschließlich ▼ Ausgabe von Ergebnissen des Dateiformats irgendein Format ▼ **③**

Datum Suche nach zuerst gesichteten Webseiten ohne Zeitbegrenzung ▼ **④**

Position Antwortseiten, in denen meine Begriffe vorkommen irgendwo auf der Seite ▼ **⑤**

© Google

Bitte kreuze in jeder Zeile die richtige Antwort an!

| | richtig | falsch |
|--|--------------------------|--------------------------|
| ① beliebige Sprache ▼ | <input type="checkbox"/> | <input type="checkbox"/> |
| ② alle Regionen ▼ | <input type="checkbox"/> | <input type="checkbox"/> |
| ③ irgendein Format ▼ | <input type="checkbox"/> | <input type="checkbox"/> |
| ④ ohne Zeitbegrenzung ▼ | <input type="checkbox"/> | <input type="checkbox"/> |
| ⑤ irgendwo auf der Seite ▼ | <input type="checkbox"/> | <input type="checkbox"/> |

Illustration 3 Exemplary test item for ICT competence (2)

Frage 20:

ITT943

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



© Microsoft

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



| | |
|--------------------------|---|
| <input type="checkbox"/> | 1 „Bearbeiten“ |
| <input type="checkbox"/> | 2 „Bild“ |
| <input type="checkbox"/> | 3  |
| <input type="checkbox"/> | 4  |

Illustration 4 Exemplary test item for ICT competence (3)

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

① Bei „Vorschau“ klicken

☐

② Bei „An Papiergröße anpassen“ klicken

☐

③ Bei „Eigenschaften“ klicken

☐

④ Bei „Folien“ klicken

☐

Illustration 5 Exemplary test item for ICT competence (4)

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

| C10 | | fx | | | | | | |
|-----|------------|----------------------|-------------------|--------------------|-------------------|---|---|---|
| | 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 € | | | |
| 8 | | | | | | | | |

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)

☐

Illustration 6 Exemplary test item reading competence

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.

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.

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 | <input type="text"/> |
| 2 | <input type="text"/> |
| 3 | <input type="text"/> |
| 4 | <input type="text"/> |
| 5 | <input type="text"/> |

- A Fruchtbares Land
- B Gräber und Schätze
- C Werkzeuge für Pyramiden
- D Wichtige Mumienfunde
- E Cleopatra
- F Götter und Könige

Illustration 7 Exemplary items for mathematical competence

MATHEMATICAL COMPETENCE OVER THE LIFESPAN

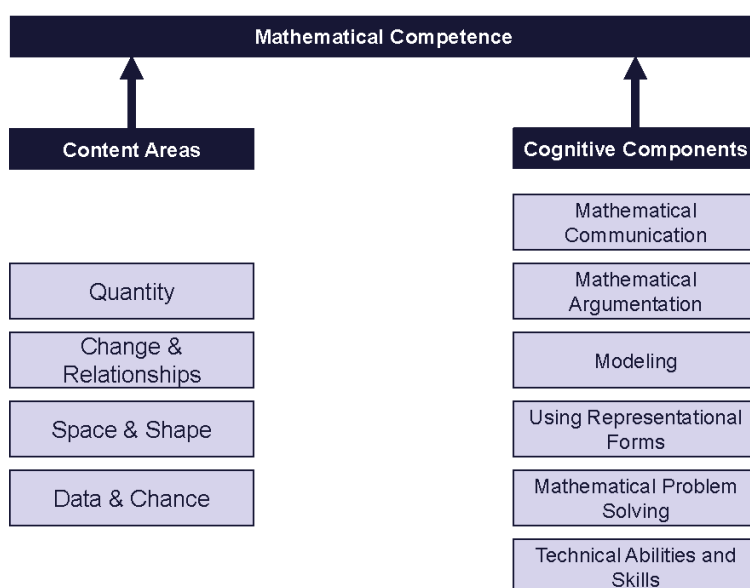


Figure 1. NEPS framework for Mathematical Competence

| Sample Item 1: Stones in the Bowl | | |
|---|---------------------------------------|--|
| <i>Age cohort(s)</i> | <i>Content area</i> | <i>Cognitive component(s)</i> |
| kindergarten – grade 1 | quantity (sets, numbers & operations) | mathematical problem solving, technical abilities and skills |
| <p>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?</p> | | |

Figure 2. Sample Item 1

| 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?

| | |
|--------------------------|-------------|
| <input type="checkbox"/> | 5 <i>m</i> |
| <input type="checkbox"/> | 8 <i>m</i> |
| <input type="checkbox"/> | 12 <i>m</i> |
| <input type="checkbox"/> | 16 <i>m</i> |

Illustration 8 Exemplary items for mathematical competence (2)**MATHEMATICAL COMPETENCE OVER THE LIFESPAN****Sample Item 3: At the Zoo**

Age cohort(s)
Grade 9 –
College
students

Content area
Change & Relationships

Cognitive component(s)
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?

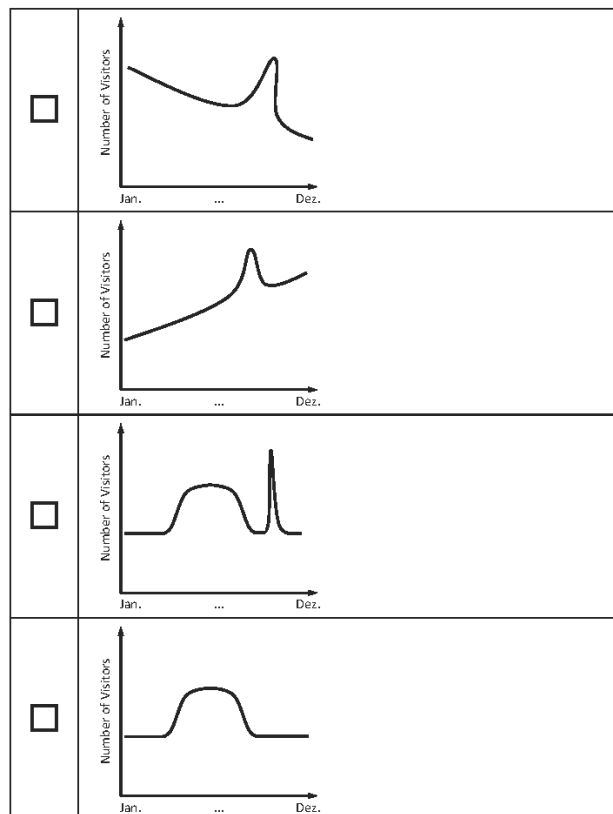


Illustration 9 Exemplary item for science competence

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!

| | |
|--------------------------|----------------|
| <input type="checkbox"/> | Versuchsplan A |
| <input type="checkbox"/> | Versuchsplan B |
| <input type="checkbox"/> | Versuchsplan C |
| <input type="checkbox"/> | Versuchsplan D |

Illustration 10 Exemplary item for declarative metacognition

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. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Er verbindet verschiedene Aufgaben miteinander. Das Geschenk kauft er auf dem Weg zum Schwimmverein. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Er lässt sich von anderen helfen. Er bittet seinen Bruder, das Geschenk zu kaufen. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Er macht nur die Hausaufgaben, die schnell gehen. Dann macht er die anderen Dinge. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Er besorgt zuerst das Geburtstagsgeschenk. Wenn das zu lange dauert, lässt er die Hausaufgaben oder das Schwimmen weg. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Er macht immer genau die Dinge, auf die er gerade am meisten Lust hat. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

12 References

- Aßmann, C., Steinhauer, H. W., Kiesel, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S. & Blossfeld, H.P. (2011). Sampling designs of the National Educational Panel Study: challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Berendes, K., Fey, D., Linberg, T., Wenz, S. E., Roßbach, H.-G., Schneider, T., Weinert, S. (2011). Kindergarten and elementary school. In: H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 203–216). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P. (1985). *Bildungsexpansion und Berufschancen*. Frankfurt: Campus.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Erikson, R. , J. H. Goldthorpe, L. Portocarero. (1979). *Intergenerational class mobility in three Western European societies: England, France and Sweden*. In: *British Journal of Sociology* 30 (1979). S. 341 – 415.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., Bos, W., & Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Ganzeboom, H. B. G. (2010). *Questions and Answers about ISEI-08*. Available from <http://home.fsw.vu.nl/hbg.ganzeboom/isco08/qa-isei-08.htm>.
- Ganzeboom, H. B. G. , de Graaf, P. M., Treiman, D. J., de Leeuw, J. (1992). *A standard international socio-economic index of occupational status*. In: *Social Science Research* 21 (1992). S. 1 –56.
- Jann, B. (2007). *fre: Stata module to display one-way frequency table*. Available from: <http://ideas.repec.org/c/boc/bocode/s456835.html>.
- Rohwer, G. (2011): *Using Sampling Weights for Model Estimation?* Available from: <http://steinhaus.stat.ruhr-uni-bochum.de/papers/dsw.pdf>.
- Treiman, D. J. (1977). *Occupational prestige in comparative perspective*. New York et al.: Academic Press.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C. (2011). *Development of competencies across the life span*. In: H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft: Special Issue 14. Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 203–216). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wegener, B. (1985). *Gibt es Sozialprestige?* In: *Zeitschrift für Soziologie* 14(3). S. 209-235, Mannheim.