

Starting Cohort 2: Kindergarten (SC2)  
SUF-Version 1.0.0  
Data Manual  
*Jan Skopek, Sebastian Pink,  
Daniel Bela*

SPONSORED BY THE



Federal Ministry  
of Education  
and Research

# Data Manual

## Starting Cohort 2

### From Kindergarten to Elementary School

(NEPS SC2 1.0.0)

Jan Skopek, Sebastian Pink, Daniel Bela  
NEPS Data Center

November 5, 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 2 – From Kindergarten to Elementary School. NEPS SC2 1.0.0. NEPS Research Data Paper, University of Bamberg.*

*This release of scientific use data from Starting Cohort 2 – “From Kindergarten to Elementary School” was prepared by the staff of the NEPS Data Center. Several experts from the whole NEPS group contributed to coding and scaling. 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 (anonymization, 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)

Judith Pfohl (data plausibility testing)

Maja-Henrieke Lomb (data plausibility testing)

Stefanie Irrler (proofreading)

**Editors of the data manual**

Jan Skopek

Sebastian Pink

Daniel Bela

National Educational Panel Study (NEPS)

Data Center

Wilhelmsplatz 3

96047 Bamberg, Germany

Contact: [userservice.neps@uni-bamberg.de](mailto:userservice.neps@uni-bamberg.de)

Web: <https://www.neps-data.de/en-us/datacenter>

Phone: +49 951 8633511

## Table of Contents

<b>1</b>	<b>Introduction .....</b>	<b>4</b>
1.1	About this manual .....	4
1.2	Obtaining the data .....	5
1.3	Three modes of data access .....	5
1.4	Respect rules of data usage! .....	6
1.5	Publications with NEPS data .....	6
<b>2</b>	<b>General Conventions.....</b>	<b>7</b>
2.1	File names.....	7
2.2	Variable names .....	8
2.3	Missing values.....	12
<b>3</b>	<b>Surveys and Sampling .....</b>	<b>15</b>
3.1	Overview .....	15
3.2	Sampling.....	16
3.3	Surveys and Tests.....	17
<b>4</b>	<b>Data Structure .....</b>	<b>20</b>
4.1	Overview .....	20
4.2	Data files.....	22
4.3	Syntax for cleaning filtered question data (PAPI) .....	26
<b>5</b>	<b>Coding.....</b>	<b>27</b>
<b>6</b>	<b>Weights .....</b>	<b>33</b>
<b>7</b>	<b>Examples .....</b>	<b>33</b>
7.1	Example 1 – First steps using CohortProfile .....	34
7.2	Example 2 – Merge datasets to CohortProfile .....	35
7.3	Example 3 – Indirectly matching different datasets .....	36
7.4	Example 4 – Prepare multilevel data.....	38
7.5	Example 5 – Using cleaning syntax for PAPI filtering .....	40
7.6	Example 6 – Using weights .....	41
<b>8</b>	<b>Rules and Recommendations .....</b>	<b>42</b>
8.1	Rules.....	42
8.2	Recommendations .....	42
<b>9</b>	<b>Tools for Stata users .....</b>	<b>43</b>
<b>10</b>	<b>Further information.....</b>	<b>45</b>
<b>11</b>	<b>Appendix .....</b>	<b>46</b>
<b>12</b>	<b>References .....</b>	<b>49</b>

# 1 Introduction

## 1.1 About this manual

This manual is intended to assist your work with the data of the *NEPS Starting Cohort 2 – From Kindergarten to elementary School* (NEPS SC2 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 2

Description	Language	Status
<i>Wenzig (2012a) &amp; Wenzig (2012b)</i> Codebook („a“ for german version, „b“ for english version)	DE/EN	available
Survey instruments package	DE/EN	available
<i>Aßmann, Koch, Steinhauer, &amp; Zinn (2012)</i> Technical report: Weighting	EN	available
<i>Pohl &amp; Carstensen (2012)</i> Technical report: Scaling the Data of the Competence Tests	EN	available
<i>Lockl (2012)</i> Competencies: Assessment of Procedural Metacognition	EN	available
<i>Koberg (2012a)</i> Technical report: Anonymization	EN	available
<i>Koberg (2012b)</i> Technical report: RegioInfas / infas geodaten	EN	available
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	available

You can download these documents here:

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

<https://www.neps-data.de/en-us/datacenter/researchdata/startingcohortkindergarten>  
(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](mailto: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](http://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](http://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 anonymization. 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 anonymized 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 “Technical report: Anonymization” (see section 1.1).

## File Format

All files are available in recent Stata and SPSS format with bilingual variable 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 for 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 form 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 2 – Kindergarten (From Kindergarten to Elementary School), doi:10.5157/NEPS:SC2:1.0.0. 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:SC2:1.0.0>

## 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]	<b>Indicator of starting cohort</b> 1 = Infants 2 = Kindergarten 3 = 5th grade students 4 = 9th grade students 5 = First-year undergraduate students 6 = Adults
[filename]	<b>Filename conventions</b> 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]	<b>Confidentiality Level</b> D = Download version R = Remote access version O = On-site version
[#]-[#]-[#] (_beta)	<b>Version</b> First digit: denotes the main release number; the main release number is incremented with every further wave release of a starting cohort, but not necessarily indicates the number of cumulated waves in a release; e.g. in starting cohort 4, the main release number 1 comprises already two wave data (first wave fall 2010 and second wave spring 2011). Second digit: indicates major updates; major updates affect the data structure (e.g. release of imputed datasets); updating your syntax files may be necessary. Third digit: indicates minor updates; minor updates affect the content of cells but not the data structure; updating your syntax files is not necessary. _beta-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 “SC2\_xTarget\_D\_1-0-0.dta” refers to the download-version of the Stata-format data file xTarget of Starting Cohort 2 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

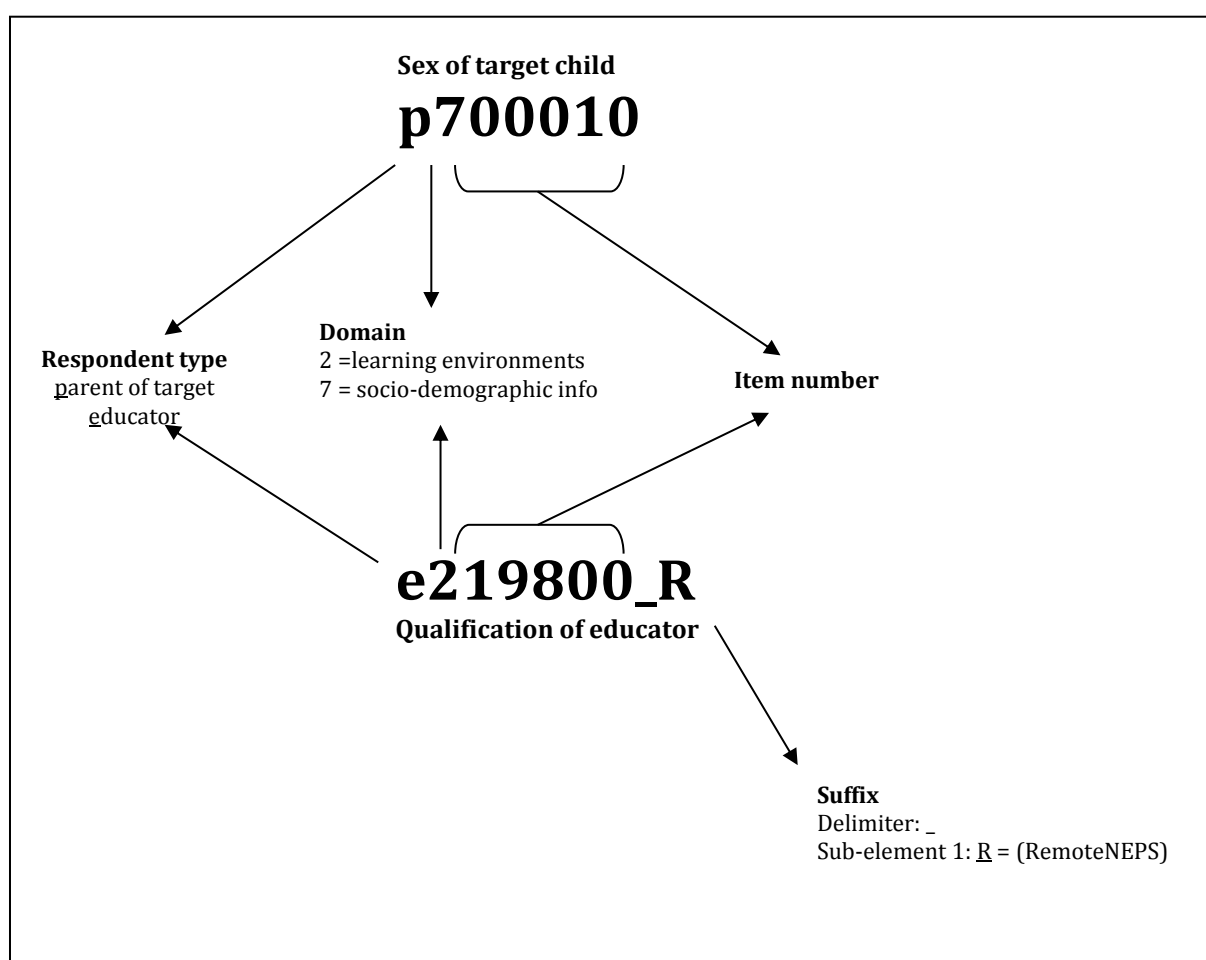


Table 3 Naming conventions of variable names

Digit	Description
1	Indicates to which <b>respondent type</b> the variable refers; in starting cohort 2, 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	<b>Topic/domain</b> (according to the theoretically coordinated dimensions of the NEPS): <ul style="list-style-type: none"> <li>1 = competence development (pillar 1)</li> <li>2 = learning environments (pillar 2)</li> <li>3 = educational decisions (pillar 3)</li> <li>4 = migration background (pillar 4)</li> <li>5 = returns to education (pillar 5)</li> <li>6 = working group “interest, self-concept and motivation”</li> <li>7 = socio-demographic information</li> <li>a = from birth to early child care (stage 1)</li> <li>b = from Kindergarten to elementary school (stage 2)</li> <li>c = from elementary school to lower secondary school (stage 3)</li> <li>d = from lower to upper secondary school (stage 4)</li> <li>e = from upper secondary school to higher edu./occup. training/labor market (stage 5)</li> <li>f = from vocational training to the labor market (stage 6)</li> <li>g = from higher education to the labor market (stage 7)</li> <li>h = adult education and lifelong learning (stage 8)</li> <li>s = basic program</li> <li>x = generated variables</li> </ul>
3–7	<b>Item number:</b> The item number typically consists of four numeric characters plus one alphanumeric character
8–11	<b>Suffix</b> (optional): Suffixes are separated from the previous characters by an underscore. There are four types of suffixes: <ul style="list-style-type: none"> <li>• 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"> <li>○ Regional codes based on the Nomenclature of Territorial Units for Statistics (NUTS) <ul style="list-style-type: none"> <li>▪ <i>g1</i> = NUTS level 1 (federal state/Bundesland)</li> <li>▪ <i>g2</i> = NUTS level 2 (government region/Regierungsbezirk)</li> <li>▪ <i>g3</i> = NUTS level 3 (district/Kreis)</li> </ul> </li> <li>○ Occupational/prestige codes <ul style="list-style-type: none"> <li>▪ <i>g1</i>: KldB 1988 (German Classification of Occupations 1988)</li> <li>▪ <i>g2</i>: KldB 2010 (German Classification of Occupations 2010)</li> <li>▪ <i>g3</i>: ISCO-88 (Internat. Standard Classification of Occupations 1988)</li> <li>▪ <i>g4</i>: ISCO-08 (Internat. Standard Classification of Occupations 2008)</li> </ul> </li> </ul> </li> </ul>

- 
- g5: ISEI-88 (Internat. Socio-Economic Index of Occupational Status 1988)
  - g6: SIOPS-88 (Standard International Occupational Prestige Scale 1988)
  - g7: MPS (Magnitude Prestige Scale)
  - g8: EGP (Erikson, Goldthorpe, and Portocarero's class categories)
  - g9: BLK (Blossfeld's Occupational Classification)
  - g10: DKZ 2010 (Documentary Code Number 2010)
  - g11: DKZ 1988 (Documentary Code Number 1988)
  - g12: Coding scheme
  - g13: KKZ (Course code / Kurskennziffer)
  - g14: ISEI-08 (Internat. Socio-Economic Index of Occupational Status 2008)
  - g15: CAMSIS (Social Interaction and Stratification Scale)
  - g16: SIOPS-08 (Standard International Occupational Prestige Scale 2008)

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

- Wide-format suffix:  
Wide-format variables stored are indicated by the suffix *\_w#* (e.g. *\_w1*, *\_w2*, etc.). Note that the wide-suffix not necessarily implies a wave logic. For instance, the presence of a set of variables *a\_w1*, *a\_w2*, ..., *a\_w10* means that there are up to 10 values for the variable "*a*" (e.g. the item corresponding to variable *a* was measured repeatedly in a questionnaire loop) relating to a row entity (e.g. a person or a school episode). Of course, there are cases where suffix *\_w#* directly relates to wave-specific values of the underlying variable.
  - Confidentiality suffix:  
This suffix pertains to all variables that were anonymized (see 1.3). The suffix indicates a variable's degree of anonymization. This suffix may either stand alone (e.g. country of birth: *t405010\_R*) or be combined with other suffixes (e.g. district of place of birth: *t700101\_g3R*).
    - O: on site; data on this variable are only available on site
    - R: remote access; data on this variable are available on site or via RemoteNEPS
    - D: download; data on this variable are available via all three modes of access
-

## 2.2.2 Special conventions for variables in test data

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

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

The SUF contains item variables (responses to the test items) as well as overall competence scores. There are two versions of item variables in the SUF: scored items named *{varname}\_c* and scored partial credit-items named *{varname}s\_c*. For example, *magD041\_c* is a scored item variable (values 0 or 1) measuring mathematical competence of children with an item targeted at grade 5 students. Note that the item variable does not necessarily indicate that the students’ mathematics skills are measured in grade 5. It could also be that the measurement was done in grade 7 and that an item was used that has already been implemented in grade 5. Additionally to the item responses, overall measures of the competence score are given. Suffix *\_sc{number}* is used for several aggregated scores and the meaning of the suffixed number is fixed as follows: 1=WLE (Weighted Maximum Likelihood estimates<sup>1</sup>), 2=standard error of WLEs, 3=sum, 4=mean, 5=difference. For example, variable *grk1\_sc3* represents the sum score of the grammar test of children being tested in the first wave (2010) in Kindergarten. Detailed descriptions on how competence scores are estimated can be found in the respective reports for the different competence domains. If there are several aggregated scores (e.g. different sum scores), letters are appended additionally (e.g. *dgg9\_sc3a* is of the sumscores for perceptual speed, while *dgg9\_sc3b* is the sumscores for reasoning – both are measures of domain general cognitive functioning).

---

<sup>1</sup> WLEs are estimated in tests that are scaled based on Item Response Theory Models.

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

Table 5 Overview of missing codes

Code	Missing
<b>Item nonresponse</b>	
-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	Anonymized
–55	Not determinable
–56	Not participated

---

We distinguish between three types of missing values:

- *Item nonresponse* occurs if a person did not respond to a question.
    - The most common instances of item nonresponse are refusals (–97) and don't knows (–98).
    - 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 p407050\_D indicating citizenship of the target child has a missing code –20 for “stateless”).
    - 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 used also for the more general case where values of a variable are not available due to design issues.
    - The second concerns individuals: If a question does not apply to a person, it is coded “Not applicable” either by the respondent's or the interviewer's remark (–93) or like it is the case for computer-assisted interviews automatically by the survey instrument (. = Filtered). In the context of paper-based questionnaires (PAPI mode) the code –99 is set for filtered variables (not by default, but after applying our filtering syntax provided together with the data, see section 4.3).
  - *Edition missings* are defined in the process of data editing.
    - Implausible values are recoded into missing (–52).
-

- Sensitive information which is only available via RemoteNEPS and/or on site access is anonymized (-53).
- Coding schemes are used to generate variables (e.g. occupational coding). If the information from the original data is not sufficient to generate a value, we assign the missing code “Not determinable” (-55).
- If a person was not present during the interview, did not fill out a questionnaire although it was administered to her, the concerning variables are assigned the missing code “Not participated” (-56). This missing code is special in so far as target persons lacking interview data (e.g. due to illness) usually are not entailed in the corresponding datasets. In the special case of one dataset integrating multiple waves widely this missing code is assigned.

**nepsmiss: Recoding missing values in Stata**

We offer a Stata ado file on our web portal which automatically recodes all missing values into extended missing values (.a, .b, etc.), and vice versa, while preserving value labels. We generally recommend running *nepsmiss* before any further data preparation. See section 9 for further information on how to install 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 Kindergarten cohort, i.e. starting cohort 2. The Kindergarten sample focuses on the population of children in year 2010/2011 attending day-care facilities (Kindergartens) two years prior to regular school enrollment in Germany. Hence, the population to which the children in the sample (also called “target persons” in the following) relate are children in Germany that are going to entry primary school in school year 2012/2013. The sample was drawn in a multi-stage approach. Generally speaking, institutions were drawn in a first step and children in a second.

One of the major goals was in assessing competences of Kindergarten children. In this regard, a picture-based testing approach has been employed. The test sessions took place at the Kindergarten facilities. In addition, day-care educators (“Erzieher”) that have been assigned to children by the Kindergarten principals (“Kindergartenleiter”) filled out a child questionnaire. To get relevant information on institutional and familial contexts of children, day-care educators of the children, principals of the Kindergarten institutions, as well as parents of the children are surveyed additionally. The Kindergarten sample will be surveyed annually and expanded at school entry. Thus, one of the major goals is capturing the children’s’ transition from Kindergarten to elementary school. For an in depth treatment of theoretical perspectives and design issues of the Kindergarten sample consult Berendes et al. (2011).

Participation in the NEPS study is voluntary. Parents of children were invited to participate in the study. Note, that contrary to starting cohort 3 and 4, in starting cohort 2 only those children were eligible targets whose parents gave a joint consent, i.e. they gave their consent for their child(ren) to participate as well as their consent to participate themselves in the parent interview.

Data collection of first wave had been projected for spring 2011. In effect, testing and questioning in the day-care facilities (NEPS study *A12*) took place from January to October 2011. Field work of *A12* was organized and conducted by the data collection institute IEA DPC (IEA Data Processing and Research Center, Hamburg). Thereby, paper-based questionnaires were administered to educators and principals of the sampled Kindergartens (PAPI mode). Parental interviews (NEPS study *B11*) were accomplished via telephone interviews (CATI mode) from April to July 2011 in a first tranche and from November to December 2011 in a second tranche. All parental interviews have been conducted by the data collection institutes infas (Institute for Applied Social Sciences, Bonn).

The released dataset contains a sample of 2,996 eligible children born in 2005 and 2006.

(p. 19) gives an overview to respondents, modes, instruments, and contents of data collection of the first wave.



## 3.2 Sampling

Testing of children and surveying of educators and principals were conducted in a nationwide, representative sample of day-care facilities. Day-care facilities were defined in accordance with the federal and state statistical offices, which relate to the operating license (§ 45 SGB VIII) and the size of the facility (at least 10 places available, from which at least five must be occupied).

As there is no complete list of Kindergarten institutions in Germany and in order to be able to accompany the four-year-old children later in the school context as well, the NEPS employed an indirect two stage approach for sampling. First, elementary schools have been sampled on a nationwide and representative basis. Schools with a grade 1 level were drawn by means of a size-proportional random selection and are at the same time the basis for the upcoming main survey in 2012 in grade 1. Since these elementary schools are to connect the early child survey with the school surveys in the National Educational Panel Study, all day-care facilities (Kindergartens) from which children may transfer on to the schools featured in this sample have been previously identified. In effect, sampled schools provided a list of linked Kindergartens and in second sampling stage a set of Kindergartens was drawn at random from this list for each school. Hence, the number of Kindergartens drawn per school depends on school size. At the sampled Kindergartens questionnaires and tests have been administered. All four-year-old children (more precisely their parents) from the Kindergartens are invited to participate. This procedure will ensure that a high number of four-year-olds in Kindergarten will also be included in the sample of first-graders.

Kindergartens and parents received flyers containing general information to the National Educational Panel Study, information on testing, as well as data privacy. Monetary incentives (50 euro in cash) in case of participation were given to Kindergartens. Children received one toy per test day.

In the first wave of the Kindergarten cohort in spring 2011, 2,996 Kindergarten children and their parents in 279 Kindergartens have agreed to participate in the panel survey. This corresponds to a response rate of 56.2%. 2,741 children and 2,340 parents as well as 831 educators and 237 principals actually participated in the first wave (cf. Table 7 on page 21).

In year 2012/2013 Kindergarten children will be integrated in the NEPS elementary school study. By sampling first grades entirely, the NEPS aims at an increased likelihood of surveying children that have already participated in Kindergarten studies. Children who do not enter into a NEPS school or start school either earlier or later will be traced individually similar to the concept of individual retracking of school cohorts.

For more information on the sampling design see Aßmann et al. (2011). Importantly, consult Supplement “Technical report: Weighting” (Aßmann et al., 2012) for weighting issues in starting cohort 2.

### 3.3 Surveys and Tests

#### 3.3.1 Assessment of competencies

Kindergarten children were tested in different domains. In first wave in 2011, they were tested in sciences, vocabulary, and grammar. Tests took place at two test days (each took 30 minutes) and were administered as individual, picture-based testing.

- Sequence on test day 1:
  - scientific competence + procedural metacognition
- Sequence on test day 2:
  - listening comprehension at sentence level: receptive grammatical competence + procedural metacognition
  - listening comprehension at word level: receptive vocabulary + procedural metacognition

In the second wave tests in mathematics, domain-general cognitive function, working memory, delay of gratification and phonological awareness will follow. Repeated testing is planned from the third wave (2013) onwards.

For an overview to the measurement of competences in Starting Cohort 2 consult the document “Information on the Competence Test” being part of Supplement “Survey instruments package”. Illustration 1 (p. 48) in the appendix shows an exemplary test item. For a detailed description on scaling of competence data read Pohl & Carstensen (2012). Further information on the assessment of procedural metacognition can be found at Lockl (2012). See Weinert et al. (2011) for a comprehensive discussion of measurement of competencies across the life span in the NEPS.

#### 3.3.2 Data on children provided by educators

With regard to questionnaire data children (target persons) were not surveyed directly. For each child the Kindergarten principal specified (at least) one educator who was designated to fill out the questionnaire “Details on the Child”. It is possible that there are several educators that filled out the questionnaire jointly.

#### 3.3.3 Data on children and home context provided by parents

To get background information about the child and the parental context, legal parents (biological or social) of the children were invited to participate in a supplementary study (NEPS study B11). Note that unlike to starting cohort 3 and 4, consent of parents and children were coupled in starting cohort 2. 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 current care arrangements have been collected.

For starting cohort 2 (Kindergarten), 3 (fifth graders) and 4 (ninth graders), a common CATI instrument has been deployed. There is a core program of questions that have 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). Table 13 in the appendix (p. 46) shows which parts of the CATI questionnaire were administered to which starting cohort's parents in the first wave.

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 (see Supplements "Field work reports from infas" and "Field work reports from IAE-DPC").

### **3.3.4 Contextual data on groups and Kindergartens**

In the educator questionnaire, educators answered questions about their own person. Additionally, the educators were also asked about some of the structural characteristics of their core group ("Stammgruppe"), if there is such a group-based organization in the Kindergarten at all, the activities within this group, and the group's composition.

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

In addition, principals filled out the so called list of children ("Kinderliste") and list of educators ("Erzieherliste"). The "Kinderliste" provides essential methodological information (e.g. needed by the NEPS methods group for generating sampling weights) on participating as well as non-participating children, however data on non-participants is not provided in the scientific use file. For participating children (i.e. parents gave consent) basic data on children (e.g. sex, date of birth, place of living, or whether German language is being spoken primarily at home) is provided from this list regardless whether educators or parents responded to questionnaires. Note that principals were invited to make guesses on those variables, if they were not absolutely sure. Nevertheless, there is a very high accordance between data given in the list and data provided by the parents (conditional on non-missing values). E.g. there is about 99% agreement on sex and year of birth and roughly 92% on the question whether German is the predominant language at home. However, when working with these variables one should remember the fact that even very similar variables base on different data-generation processes characterized by different modes and respondents (data provided by principals, educators and parents).

### **3.3.5 Logic of assignment of children and educators to Kindergarten groups**

Usually, child care in Kindergarten is organized in groups being supervised by one or more educators. Educators can be responsible for one or more groups. Sometimes, there are Kindergartens working with "open" groups or no groups at all. In this case, the whole Kindergarten can be considered as one group. Consequently, (almost) every child is uniquely assigned to one group (at one wave). However, while for the most part educators are uniquely assigned to one group, there are educators that care for two or more groups. Consequently, there is an n:m relationship between groups and educators

that must be considered carefully when preparing the data. Further treatment of this issue as well as practical advices will be given in section 4.2.2 and 7.3 below.

Table 6 Overview of Respondents, Instruments, Modes and Contents of the first wave (2011) of Starting Cohort 2.

Respondent	Mode (Instrument-ID)	NEPS Unit	Contents	
Target Person (child)	Picture-based Tests	pillar 1 stage 2	sciences, vocabulary Grammar	
Parent of child	CATI (31)	pillar 2 pillar 3 pillar 4 pillar 5 AG ISM <sup>2)</sup> stage 2	HLE <sup>1)</sup> , home activities, satisfaction with Kindergarten cultural capital language use and proficiency Health TASB, SDQ current care arrangements, early school enrolment, language support	
Educator	PAPI (22)	<b>Educator's questionnaire: questions about the group</b>		
		pillar 2	structural characteristics, equipment and activities	
		pillar 3	social composition	
		pillar 4	ethnic composition, German-language use	
		<b>Educator's questionnaire: personal questions<sup>3)</sup></b>		
		pillar 2	educator's pedagogic orientations and vocational background	
		pillar 3	Sociodemographics	
		pillar 4	migration background, L1, language use	
	PAPI (28)	<b>Child questionnaire: questions about child</b>		
		pillar 2	children's activities, day of school enrollment	
		pillar 4	German-language proficiency of migrants	
		AG ISM <sup>2)</sup>	SDQ, TASB	
Principal	PAPI (29)	<b>questions about the institution</b>		
		pillar 2	structural characteristics, composition, equipment, pedagogic orientations, composition of staff	
		pillar 3	social composition	
		pillar 4	ethnic composition, availability and German-language use and L1 support programs	
		pillar 5	Competition	
		stage 2	type of language skills assessment, language support	
		<b>personal questions<sup>3)</sup></b>		
		pillar 2	vocational background of the head	
		pillar 3	Sociodemographics	
		pillar 4	migration background of the educator	
		list of children (paper based)	"Kinderliste": Basic sociodemographics to children (participants & non-participants), assignment to group, ID of educator(s) filling out the child questionnaire	
		list of educators (paper based)	"Erzieherliste": Assignment of educators to groups	

<sup>1)</sup> "H" is used for home; „LE“ is used for learning environment. <sup>2)</sup> NEPS workgroup Interest, Self-Concept, Motivation. <sup>3)</sup> In second wave these questions will only be asked if the educator/principal has changed.

## 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 an 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 is, 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 provided so called spell files that are prefixed by a "sp". An example in starting cohort 2 is the file *spChildCare* that contains a target's child care 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 2.

Note, that since Starting Cohort 2 has 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 child). ID\_t is unique over waves and over starting cohorts.
- **ID\_e:** Identifies an educator uniquely. It can be used to match educator data to children observations via Kindergarten groups. ID\_e is unique over waves and over starting cohorts.
- **ID\_p:** Identifies a 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 children having parents, who gave their consent but could not be contacted in the data collection period (e.g. due to unidentifiable address data) and thus are not included in the parent files, there is a missing code of -55 ("Not determinable").
- **ID\_group:** Identifies a Kindergarten group uniquely only within a wave. It will not be unique over waves. Note, there are children who lack an assignment to any group (for unknown reasons). For those cases ID\_group is coded with the missing code -90 ("Unspecific missing").
- **ID\_i:** Identifies a Kindergarten 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 2 (Kindergarten)

File	Unique Identifier	Content	N	Availability*
<b>Para data and linkage</b>				
CohortProfile	ID_t	Para data on all children of the sample.	2996	D,R,O
Groups	ID_group ID_e	Assignment of classes (Kindergarten groups) to educators including data on groups collected in the educators questionnaire	720	D,R,O
ParentMethods	ID_p	Para-data from the parent CATI. Includes all parents that could be contacted in the field.	2845	D,R,O
<b>Survey and test data collected at Kindergarten</b>				
xTarget	ID_t	Data to Kindergarten child provided by educators	2741	D,R,O
xTargetCompetencies	ID_t	Competence test data on children	2949	D,R,O
xEducator	ID_e	Data from Educators	831	D,R,O
xInstitution	ID_i	Data from the principals of Kindergartens	237	R,O
<b>Data from parents</b>				
xParent	ID_p	Most of data from parent's CATI	2340	D,R,O
spChildCare	ID_p sptype spell	Child care spells. Child care history collected in the parent interview, but stored in spell format	3889	D,R,O
<b>Regional data</b>				
xInstitutionRegioInfas	ID_i regio	Regional data on the institution's address	1116	O
xTargetRegioInfas	ID_t regio	Regional data on the child's parents' address	11380	O

## 4.2 Data files

### 4.2.1 CohortProfile: Para data on the cohort's panel sample

The file *CohortProfile* contains all target persons of the panel sample. In SC2 these are all children whose parents agreed to their child's and their own participation in the NEPS survey. The file has a long format structure with variables *ID\_t* and *wave* identifying a row uniquely (see Figure 2 for a data example). For each wave and child it contains information on sample assignment, data availability and participation status. Furthermore, data from the list of children ("Kinderliste", filled out by the principals) are stored in this file. This is valuable data since due to nonresponse only for a subset of children data from parental interviews is available. In the example below, there are three children born in 2005 and two in 2006 (variable *tx8050y*), three of them are male and two are female (variable *tx80501*) and for all of them there is data available on the Kindergarten level (variable *tx80524*). Additionally, weighting variables are stored in this file. In general, we strongly recommend using this file as a starting point of any analysis (see syntax examples in section 7).

Figure 2 Data example of data file *CohortProfile*

<i>ID_t</i>	<i>wave</i>	<i>ID_p</i>	<i>ID_i</i>	<i>ID_group</i>	<i>tx80524</i>	<i>tx80501</i>	<i>tx8050y</i>
2000562	1	1021206	1000859	1000859104	1	2	2005
2000563	1	1021125	1000824	1000824102	1	1	2005
2000564	1	1021000	1000816	1000816101	0	1	2006
2000565	1	1021010	1000874	1000874101	1	1	2006
2000566	1	1020932	1000824	1000824102	1	2	2005

### 4.2.2 Groups: Mapping of educators and groups

Usually, child care in Kindergarten is organized in groups being supervised by one or more educators, while educators also can work in several groups. Generated on the basis of the list of educators (see section 3.3.4, p. 19) the file *Groups* is dedicated to depict this rather complex n:m-relationship between educators and groups. In particular, the file is essentially needed for preparing data on the very group context where children within a Kindergarten are exposed to.

*Groups* provides educator-group ID assignments (variables *ID\_group* and *ID\_e*) in a simple long format (for a data example see Figure 3). Thereby, one row represents uniquely one group-educator assignment (per wave). Naturally, there are two perspectives on these data. You might be interested in educators and might want find out the groups they are supervising. Alternatively, but maybe more importantly, you are interested in groups. Thus, for a given group you want to assess the number and IDs of all the educators that are assigned to this group (see syntax example in section 7.3, p. 36). Both pathways are very easy to accomplish using the *Groups* file.

**Figure 3 Data example of data file *Groups***

ID_group	wave	ID_e	ex20100
1000687101	1	1004373	1
1000687101	1	1007517	0
1000689101	1	1003945	1
1000691101	1	1003753	1

For providing a convenient 1:n relationship, i.e. only one educator per group, we deliver a dummy variable (*ex20100*) indicating a default group-educator assignment. This default assignment is derived according to an abstract data edition rule: The educator validly answering most questions about the group is determined as default, and flagged with the value 1. If such a person cannot be determined, a default educator is drawn randomly. It should be appropriate to use this default educator for most analyses.

#### **4.2.3 xTarget: Educator's answers on children**

In general, children are the target persons in the Kindergarten and school cohorts of the NEPS. Within the context of the Kindergartens, questionnaire data on children was collected via a PAPI questionnaire. Note that in Starting Cohort 2 those questionnaires were administered to and filled out by the educators as proxies of the children. The coded data is stored in the file *xTarget* following a simple cross-sectional structure (i.e. 1 row = 1 child). Children are identified by ID\_t. Since educators were the respondents variable names begin with “e” which follows the general naming convention (see section 0). More than one educator could be eligible for answering the child questionnaire, however mostly only one educator was specified. The IDs of those educators, who were invited to fill out the child questionnaire, are delivered in *xTarget*, too. Note that these IDs relate only to those educators, who were invited to fill out the child's questionnaire, nothing more. In particular, they tell us nothing neither about the exact educator, who finally filled out the questionnaire, nor necessarily the educator who is supervising a child's group. For the latter information consult the *Groups* file.

Note that due to nonresponse on the side of the educators data of the proxy questionnaire is not necessarily available for all children. Consequently, not all children of the panel sample (as depicted by the file *CohortProfile*) have entries in *xTarget* (about 8.5% have no entries). You can check this very easily by tabulating the variable *tx80521* in file *CohortProfile*, which contains the whole sample.

Since already collected by the list of children some basic information on children (like sex and date of birth) are not surveyed again in the children questionnaires administered to the educators. To enhance usability of the file, those variables (tx805..) were copied from *CohortProfile* to *xTarget*.

#### **4.2.4 xTargetCompetencies: Test data of children**

Three different competence tests were administered directly on the children (sciences, vocabulary, grammar). The scored results of these tests can be found in file



*xTargetCompetencies* (cross-sectional format, i.e. 1 row = 1 child). In 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). Variables in competence data files follow a nomenclature that is slightly different to other files (refer to section 2.2.2 for details).

#### **4.2.5 xEducator: Data on educators and their core groups**

File *xEducator* provides data from the educator's questionnaire in a simple cross-sectional format (1 row = 1 educator). An educator is identified in the file via *ID\_e*. Note that this file does not comprise every educator somehow participating in data collection. For instance, there are some educators who only answered a child questionnaire, but not the educator's questionnaire.

#### **4.2.6 xInstitution: Answers from principals**

Answers to the principal questionnaire provided by the Kindergarten principals are stored in *xInstitutions*. 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 SC2 scientific use file.

#### **4.2.7 Data from the parent interview**

Parents' data comprise the actual data generated by the CATI questionnaire as well as method data generated in the context of the interview (e.g. para data 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 parents' interviews in several data files. Table 14 in the appendix (p. 47) 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 2.

##### *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*.

##### *spChildCare*

Specifically for starting cohort 2, the target child's caring history is surveyed retrospectively. Since this data follows an episode structure it is stored in a separate file named *spChildCare*. Care history for six types of care (like care by relatives, by au-pair, or by nanny) has been surveyed. Types are identified through the variable *sptype*. For each type all caring episodes are collected retrospectively. Each episode has a number

(variable *spell*). Therefore, data file *spChildCare* has a long data structure with one row marking one caring episode of a specific type reported by a specific parent. Thus, a row is uniquely identified (within a wave) by *ID\_p*, *sptype*, and *spell*. The enumerator variable *spell* identifies a care spell within a specific care spell type (*sptype*) reported by the parent (*ID\_p*). In the example below there are the child care histories of two children displayed (reported by their parents). While the first child had three caring episodes with different types, the second child had also three different types of episodes but two times an episode of type 1.

Figure 4 Data example of data file *spChildCare*

ID_p	sptype	spell	startm	starty	endm	endy
1004292	1	1	8	2007	4	2011
1004292	5	1	6	2006	8	2007
1004292	6	1	6	2006	8	2007
1004298	1	1	2	2007	2	2008
1004298	1	2	2	2009	4	2011
1004298	4	1	2	2008	2	2009
1004298	6	1	9	2005	4	2011

### *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 identified by *ID\_p* but also target identifier *ID\_t* is included for mapping parents to targets. For a subset of those cases an interview was effectively realized in first wave. Thus, *ParentMethods* includes more cases than *xParent*. 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.<sup>2</sup> 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.

<sup>2</sup> Nevertheless they are included in the file since they gave their consent prior at the time of invitation.

**Figure 5 Data example of data file *ParentMethods***

ID_p	ID_t	wave	intm	inty	px80220
1004292	2000905	1	4	2011	1
1004297	2000949	1	4	2011	1
1004313	2002225	1	4	2011	1
1020203	2002869	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: *RegioInfas* (generated files)

Fine-grained regional data is provided in the data files *xTargetRegioInfas* and *xInstitutionRegioInfas*. These files have been generated from the *infas geodaten* database.<sup>3</sup> Both comprises geographical information on four regional levels (coded in variable *regio*): municipality, postal code, quarters (living areas), and street sections. These data were linked to 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 Koberg (2012b).

### 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.

---

<sup>3</sup> This database is provided by the *infas geodaten* GmbH, see: <http://www.infas-geodaten.de>

The cleaning syntax is part of Supplement “Syntax package” and can be downloaded from the SC2 data section on our web portal (see section 10, p. 45). 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 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.

Table 8 Overview of coded variables

Classification	Included in	Description
KldB88	<i>xEducator; xInstitution; xParent</i>	German Classification of Occupations 1988 (4-digit)
KldB2010	<i>xEducator; xInstitution; xParent</i>	German Classification of Occupations 2010 (5-digit)
ISCO-88	<i>xEducator; xInstitution; xParent</i>	International Standard Classification of Occupations 1988 (4-digit)
ISCO-08	<i>xEducator; xInstitution; xParent</i>	International Standard Classification of Occupations 2008(4-digit)
BLK	<i>xEducator; xInstitution; xParent</i>	Occupational classification by Blossfeld based on KldB92 (cf. Blossfeld 1985; Schimpl-Neimanns 2003)
ISEI-88	<i>xEducator; xInstitution; 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; xInstitution; 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; xInstitution; xParent</i>	Metric scale to measure prestige of occupations based on ISCO-88 (cf. Treiman 1977)
SIOPS-08	<i>xEducator; xInstitution; xParent</i>	Metric scale to measure prestige of occupations based on ISCO-08
MPS	<i>xEducator; xInstitution; xParent</i>	Magnitude prestige score of occupations (cf. Wegener 1985)
EGP	<i>xParent</i>	Class scheme which assigns occupations to classes (Erikson et al. 1979)
CAMSIS	<i>xParent</i>	Classification to measure social interaction and stratification (Prandy 2000)
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 (UNESCO 2006)
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]	Selfemployed with employees	Selbständige mit unterstellten Mitarbeitern
6	[IVb]	Selfemployed no employees	Selbständige ohne unterstellte Mitarbeiter
7	[IVc]	Selfemployed Farmer	Selbständige in der Landwirtschaft
8	[V]	Manual Supervisors	Arbeiter, Techniker, Facharbeiter
9	[VI]	Skilled Worker	Qualifizierte Arbeiter
10	[VIIa]	Unskilled Worker	Unqualifizierte Arbeiter
11	[VIIb]	Farm Labor	Landwirte

Table 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]	Semiprofessions	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 kindergarten children (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 labeled as *weight\_design*. Note that also standardized weights with mean one are provided, which are often used in statistical analysis. These are named as *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.<sup>4</sup> In future releases of this manual we will extend the examples by code in R and SPSS.

---

<sup>4</sup> 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 children. 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

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

* load CohortProfile
use "SC2_CohortProfile_${version}", clear

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

* Check status of survey participation
* (> 99% participation, here defined as
*   having participated in tests or
*   having target data provided by the educators)
fre tx80220

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

* check data availability: test data and parent data
* (~77% of the sample has both test and parent data)
tabulate tx80522 tx80523, cell

```

## 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's information provided by his or her educator and (b) data provided by the child'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

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

* load CohortProfile as master dataset
use "SC2_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 "SC2_xTarget_${version}", nogen keep(master matched)

* merge file xParent (consider missing parent IDs for some children)
nepsmiss 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 "SC2_xParent_${version}", nogen keep(master matched)

```

### 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 child-level outcome variable conditional on some variation on the level of Kindergarten groups. For simplicity, say group-level information can be measured by characteristics of the very educators who take care of the respective groups.

In a first step you start with the *CohortProfile* and merge the child's information entailed in *xTarget*. As *CohortProfile* provides the child's group affiliation, you can easily merge the group's characteristics – after you have generated them according to your specific research question. To do so, you have to start with the dataset *Groups*. This dataset reflects the n:m relationship between educators and groups: One group can consist of multiple educators and one educator can be related to multiple groups. By merging the educator's information contained in *xEducator* to *Groups* 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 2. Everything you have to do after this procedure is to aggregate the information provided by the educators according to a rule. In this example we use simple means.

### Example 3 in Stata

```

/*****
* Example 3: Indirectly match educators to children
              by preparing group-level data
Procedure
1. Prepare group-level data: "sex-ratio of educators"
2. Load CohortProfile
3. Merge prepared group-level data
4. Analyse data

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

* 1) Prepare group-level data: "sex-ratio of educators"
* load Groups file
use "SC2_Groups_${version}", clear

* merge sex of assigned educators from xEducator
merge n:1 ID_e using "SC2_xEducator_${version}", ///
              nogen keep(master matched) keepusing(e761110)
nepsmiss e761110

* generate a dummy indicating female educators
recode e761110 (1 = 0) (2 = 1), gen(edufem)
drop e761110

* reorganise dataset: reshape to wide over educators
bysort ID_group (ID_e): generate number = _n
keep ID_group number edufem
reshape wide edufem, i(ID_group) j(number)

* generate group variable "sex ratio of educators"
egen sexratio = rowmean(edufem*)
drop edufem*

* temporarily save prepared group file
tempfile groupcontext
save `groupcontext', replace

* 2) Load CohortProfile file
use ID_t ID_group wave using "SC2_CohortProfile_${version}", clear

* 3) Merge group-level data
nepsmiss ID_group
merge n:1 ID_group using `groupcontext', nogen keep(master matched)

* 4) Analyse data (e.g. inspect distribution of group's sex ratio over children)
summarize sexratio, detail

```

## 7.4 Example 4 – Prepare multilevel data

In the following example, we will choose a multilevel approach for analyzing competencies of children in Kindergartens. 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—child, group, and Kindergarten level—and from multiple sources—*xTarget*, *xParent*, *xTargetCompetencies*—can be matched together for serious analyses.

### Example 4 in Stata

```

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

Goal:
Regress competence data on traits of child,
traits of Kindergarten group, and
traits of Kindergarten institutions.
Account for nested data structure.

1) prepare group level data (sex and age of educator)
2) prepare data from parental context (highest ISEI)
3) prepare institution level data (number of children)
4) prepare competence data (sum scores of vocabulary)
5) prepare proxy data on child provided by educator
6) integrate data using CohortProfile
7) analyze data (linear mixed model)

Note: replace ${version} to your file version, e.g. 'R_1-0-0'
*****/
* 1) prepare group-level data (resolve n:m relationship between groups and
educators)
* load Groups file
use "SC2_Groups_${version}", clear
keep ID_group ID_e
* keep first observed educator per group as proxy (arbitrary decision!)
bysort ID_group (ID_e): keep if _n==1
* merge proxy educators sex and age (classified)
merge n:1 ID_e using "SC2_xEducator_${version}", nogen keep(master matched) ///
      keepusing(e761110 e76112y_D)
* recode missings
nepsmiss _all
generate g_edmale = e761110==1 if !missing(e761110)
*recode categories of e76112y_D because number of cases in category 1 is not
sufficient for analysis
recode e76112y_D (1/2=1), generate(g_edage)
* temporarily save
tempfile groupcontext
save `groupcontext', replace

* 2) prepare data from parents (highest ISEI of parental context)
use "SC2_xParent_${version}", clear
nepsmiss 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 data from the institutional level (number of children)

```

```

use "SC2_xInstitution_${version}", clear
nepsmiss _all
* generate variable measuring number of children in a Kindergarten
generate kg_nofc = h217001+h217002
keep ID_i kg_nofc
* temporarily save
tempfile kgcontext
save `kgcontext', replace

* 4) prepare competence data (simplification: sum scores of vocabulary)
use "SC2_xTargetCompetencies_${version}", clear
* standardize sum score of vocabulary test
nepsmiss vok1_sc3
egen zvoc = std(vok1_sc3)
keep ID_t zvoc
tempfile zvoc
save `zvoc', replace

* 5) prepare proxy data on child provided by educator
* (scales for pro-social and problematic behavior)
use "SC2_xTarget_${version}", clear
keep ID_t e67801a_g1 e67801c_g1
nepsmiss _all
rename e67801a_g1 prosoc
rename e67801c_g1 problem
tempfile child
save `child', replace

* 6) integrate data using CohortProfile
use "SC2_CohortProfile_${version}", clear
nepsmiss _all
merge n:1 ID_group using `groupcontext', keep(master matched) nogen
merge n:1 ID_p using `parentcontext', keep(master matched) nogen
merge n:1 ID_i using `kgcontext', keep(master matched) nogen
merge 1:1 ID_t using `child', keep(master matched) nogen
merge 1:1 ID_t using `zvoc', keep(master matched) nogen

* generate further variables for analysis
generate male = tx80501==1 if !missing(tx80501)
generate ybirth = tx8050y if !missing(tx8050y)
generate lkg_nofc = ln(kg_nofc)
* keep relevant variables
keep ID_t ID_group ID_i g_edmale g_edage lkg_nofc phisei08 ///
    male ybirth prosoc problem zvoc

* 7) analyse data (linear mixed model)
* do a complete case analysis, listwise deletion of missings
egen nofmiss = rowmiss(ID_t ID_group ID_i g_edmale g_edage lkg_nofc phisei08 male
ybirth prosoc problem zvoc)
drop if nofmiss > 0
* estimate a linear mixed model that explains variation in zvoc by
* child-, group-, and institution-level variables.
* define a random intercept model with 2 random effects,
* one at the level of institution and one at the level of groups
xtmixed zvoc male ybirth phisei08 prosoc problem ///
    i.g_edage g_edmale lkg_nofc || ID_i: || ID_group:
* 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 Kindergarten principals. 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 xInstitution  
use "SC2_xInstitution_${version}", clear  
  
* e.g. variable hb1005c  
* Before: missing codes -95 (Implausible) and  
*         -90 (Unspecific missing)  
fre hb1005c  
  
* apply filter syntax  
* (assuming do-file stored in the same directory like the data file)  
quietly: do "SC2_xInstitution_${version}.do"  
  
* e.g. variable hb1005c  
* After: missing codes -95 (Implausible),  
*        -90 (Unspecific missing), and  
*        -99 (Filtered)  
fre hb1005c  
  
* 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 *e67801a\_g1* (SDQ-scale: sum prosocial behavior) 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 "SC2_CohortProfile_${version}", clear

* merge SDQ-scale variable prosocial behaviour from xTarget
merge 1:1 ID_t using "SC2_xTarget_${version}", assert(match master) keep(match)
keepusing(e67801a_g1) nolabel nogenerate

* encode missings
nepsmis e67801a_g1 tx80501

* calculate means without weighting
mean e67801a_g1, over(tx80501)

* calculate means using probability weights
mean e67801a_g1 [pweight=weight_design_std], over(tx80501)

* calculate means using sampling weights (svy-command)
svyset ID_i [pweight=weight_design_std]
svy: mean e67801a_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 you are not allowed to publish any analyses that aim for or allow a direct comparison of the German Bundesländer. Any form of “rankings” of German Länder using the NEPS data is strongly prohibited.
- keep secret the NEPS data provided!
- keep secret transmitted access codes (e.g. individual identification and password)!
- refrain from any action aimed at and suitable for re-identifying persons, households or institutions (e.g. education or support facilities)!
- refrain from mixing the data, and neither partially, with other data permitting the reidentification of persons!
- immediately inform NEPS of any accidental reidentification 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 data center (see section 10). Also, if you encounter any security leaks regarding data protection and data security, or any data quality deficiencies please inform the NEPS data center (see section 10).

### 8.2 Recommendations

We strongly recommend you to examine the data critically when you work with this release. While the NEPS invested a lot to ensure the integrity of the provided data, the latter 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!
- check documentation material and survey instruments that can be downloaded on the NEPS website (see section 10)!
- if you encounter problems or even errors in the data please contact the data center of the NEPS (see section 10)!

## 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/startkohortekindergarten>  
(in German)

<https://www.neps-data.de/en-us/datacenter/researchdata/startingcohortkindergarten>  
(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](mailto: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

Table 13 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 14 Datasets containing data from the parent interview by starting cohort (1st wave)

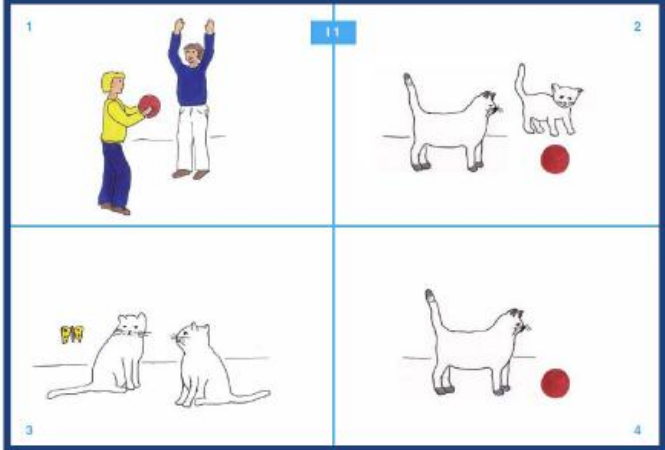
Data files	Unique Identifiers	Content	Starting Cohort		
			2	3	4
xParent	ID_p	Most data from CATI interview stored in cross-sectional format (1 row = 1 responding parent)	x	x	x
spChildCare	ID_p sptype spell	History of child care for target child in spell format (1 row = 1 care spell)	x		
spSchool	ID_p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



Illustration 1 Exemplary item from the TROG-D test of measuring grammar comprehension (translated to English, for original source see Fox 2006).

„the cats are looking at the ball“

1112



**Distractors:**

- the cat is looking at the ball
- the cats are looking at the butterfly
- the boys are playing with the ball

## 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.
- Aßmann, C., Koch, S., Steinhauer, H. W., & Zinn, S. (2012). Starting Cohort 2: Kindergarten (SC2), SUF-Version 1.0.0, Data Manual [Supplement]: Weighting. NEPS Research Data Papers, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2\\_1-0-0\\_Weighting\\_EN.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_1-0-0_Weighting_EN.pdf).
- 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, London.
- Fox, A. V. (2006). TROG-D. Test zur Überprüfung des Grammatikverständnisses. Handbuch. Das Gesundheitsforum. Idstein: Schulz-Kirchner Verlag.
- Ganzeboom, H. B. G. (2010). Questions and Answers about ISEI-08. Available at: <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 at: <http://ideas.repec.org/c/boc/bocode/s456835.html>.
- Pohl, S. & Carstensen, C. H. (2012). NEPS Technical Report – Scaling the Data of the Competence Tests. NEPS Working Paper No. 14, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/Working%20Papers/WP\\_XIV.pdf](https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf).
- Koberg, T. (2012a). Starting Cohort 2: Kindergarten (SC2), SUF-Version 1.0.0, Data Manual [Supplement]: Anonymisation. NEPS Research Data Paper, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2\\_1-0-0\\_Anonymisation\\_EN.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_1-0-0_Anonymisation_EN.pdf).
- Koberg, T. (2012b). Starting Cohorts 2, 3, 4, and 6, Data Manual [Supplement], RegioInfas (infas geodaten), v 1.0 September 21, 2012. NEPS Research Data Paper, University of

- Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC\\_RegioInfas\\_1-0\\_EN.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC_RegioInfas_1-0_EN.pdf).
- Lockl, K.* (2012). Starting Cohort 2: Kindergarten (SC2), Starting Cohort 3: 5th Grade (SC3), Starting Cohort 4: 9th Grade (SC4), SUF Version 1.0.0, Competencies: Assessment of Procedural Metacognition. NEPS Research Data Paper, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2\\_SC3\\_SC4\\_1-0-0\\_com\\_mp.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_SC3_SC4_1-0-0_com_mp.pdf).
- Prandy, K.* (2000). The social interaction approach to the measurement and analysis of social stratification. In: *International Journal of Sociology and Social Policy* 19(9/10/11): 204-236.
- Rohwer, G.* (2011): Using Sampling Weights for Model Estimation? Available at: <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.
- UNESCO* (2006). International Standard Classification of Education ISCED 1997.
- 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.
- Wenzig, K.* (2012a). Startkohorte 2: Kindergarten (SC2), SUF-Version 1.0.0, Data Manual [Supplement]: Codebook (de). NEPS Research Data Paper, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2\\_1-0-0\\_Codebook\\_de.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_1-0-0_Codebook_de.pdf).
- Wenzig, K.* (2012b). Startkohorte 2: Kindergarten (SC2), SUF-Version 1.0.0, Data Manual [Supplement]: Codebook (en). NEPS Research Data Paper, University of Bamberg. Available at: [https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2\\_1-0-0\\_Codebook\\_en.pdf](https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_1-0-0_Codebook_en.pdf).