NEPS
National Educational Panel Study

Research Data

Curricular Reform Study in Thuringia (TH)
SUF Version 1.0.0
Data Manual

Markus Zielonka, Jan Skopek, Marcel Raab

SPONSORED BY THE

Federal Ministry of Education and Research

Data Manual

NEPS – Additional Study

Organizational Reform Study in Thuringia

(NEPS TH 1.0.0)

Markus Zielonka, Jan Skopek, Marcel Raab
NEPS Data Center

November 5, 2012

Research Data Papers
at the NEPS Data Center, Bamberg

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

Citation of the manual:

Raab, M, Skopek, J and M. Zielonka (2012). Data Manual. Organizational Reform Study in Thuringia. NEPS TH 1.0.0. NEPS Research Data Paper, University of Bamberg.

This release of scientific use data from the *NEPS Additional Study I – "Organizational Reform Study in Thuringia" was prepared by the staff of the NEPS Data Center in tight collaboration with colleagues from the NEPS-Methods Group (weighting) and the staff of NEPS Pillar 1 and NEPS Stage 5 (for scoring and scaling of competencies). It represents a major collective effort. Most notably, over 15.000 lines of code and almost 130 revisions, in approximately 200 days of work, were produced in the process of data preparation and editing. The contribution of the following staff members of the NEPS is gratefully acknowledged:*

Data preparation, editing, and scaling
Daniel Bela (integration of metadata, data preparation tools in Stata)
Christoph Duchhardt (scoring and scaling of Math competencies)
Tobias Koberg (anonymisation, regional data, translation)
Manuel Munz (coding and classification)
Marcel Raab (file integration)
Benno Schönberger (weighting)
Jan Skopek (testing)
Wolfgang Wagner (scoring and scaling of all other competencies)
Knut Wenzig (management and editing of metadata, documentation)
Markus Zielonka (edition and reintegration of data)

Data manual
Markus Zielonka
Jan Skopek
Marcel Raab

National Educational Panel Study (NEPS)
Data Center
Wilhelmsplatz 3
96047 Bamberg, Germany
Contact: userservice.neps@uni-bamberg.de
Web: **https://portal.neps-data.de/de-de/datenzentrum.aspx**
96047 Bamberg, Germany
Contact: userservice.neps@uni-bamberg.de
Web: https://portal.neps-data.de/de-de/datenzentrum.aspx

Table of Contents

1 Introduction

1.1 About this manual

This manual is intended to assist your work with the data of the NEPS additional study "*Organizational Reform Study in Thuringia"* (NEPS TH 1.0.0). We aim at providing a guide of how to use these data for your research. Therefore, our focus is on practical aspects of data usage such as the dataset structure, key variables, and examples of data retrievals.

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 reference supplementary documents presenting additional information that we consider essential for working with our data:

- Codebook
- Technical reports/working papers on:
  - Weighting (Schönberger & Aßmann 2012)
  - Anonymization (Koberg 2012)
  - Scaling of Math competencies (Duchhardt, forthcoming)
  - Scaling of Physics, Biology and English competencies (Wagner, forthcoming)

You can download these documents here:

https://www.neps-data.de/de-de/datenzentrum/forschungsdaten/zusatzstudiethüringen
(german)

https://www.neps-data.de/en-us/datacenter/researchdata/additionalstudythuringia
(english)

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

userservice.neps@uni-bamberg.de

1.2 Obtaining the data

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

- Sign the data use contract and mail it to us. Click here for instructions:
    - For German users: https://www.neps-data.de/de-de/datenzentrum/datenzugangswege/datennutzungsverträge
    - For non-German users: https://www.neps-data.de/en-us/datacenter/dataaccess/datauseagreements
- After approval, sign in as a registered NEPS user at the login at www.neps-data.de
- Access the data via one of our three access modes (see below)

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

1.3 Three modes of data access

We offer you three modes of access to the data:

- Download from our website,
- RemoteNEPS (remote access via a virtual desktop),
- 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 anynomization. These data are available to registered users from the web portal via a secure connection. Files offered via RemoteNEPS contain more sensitive data within a controlled environment. The analysis of information in high resolution (e.g., fine-grained regional information) is only provided on-site in Bamberg where these data are available within a secure site. For details on the access modes, see our website at

https://www.neps-data.de/de-de/datenzentrum/datenzugangswege
(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. That is, any data contained within a less sensitive level is also included in the higher level(s).

An overview on which types of data are offered at each of these levels as well as detailed information on how the "on-site", "remote access" and "download" versions of the data were generated can be found in the "Technical Report on Anonymisation" (Koberg 2012).

File Format

All files are available in Stata and SPSS format with bilingual variable labels and value labels (German and English). Data stored in Stata format contain both languages within one file (see section 7). SPSS files are delivered separately in both languages.

1.4 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): Organizational Reform Study in Thuringia. The NEPS data collection is part of the Framework Program 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:TH:1.0.0

2 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 1 Naming conventions of file names

Element	Definition
TH	**Indicator of "Organizational Reform Study in Thuringia"**
[filename]	**Filename conventions** Prefix: x = pooled cross sectional file Keyword/mnemonic: A keyword or mnemonic indicates the content of the corresponding file (e.g., xCourse contains data from the course-teacher questionnaire) Filenames of generated datasets do not have a prefix and always start with a capital letter (e.g., *Profile*)
[D,R,O]	**Confidentiality Level** D = Download version R = Remote access version O = On-site version
[#]-[#]-[#] (_beta)	**Version** First digit: denotes the main release number; since the data from the organizational reform study in Thuringia will be released for both cross-sections at once and integrated into the same file in long format, the main release number will not change. 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 "TH_xTarget_D_1-0-0.dta" refers to the download-version of the Stata-format data file xTarget of the "NEPS - Organizational Reform Study in Thuringia" of data release 1.0.0.

2.2 Variable names

The organizational reform study in Thuringia contains data of two cross-sectional surveys and tests in one federal state and focusses on a very specific institutional change in Germany. Hence, contrary to the common NEPS naming convention of variables we sometimes provide variable names derived from German abbreviations of the questions or the items in focus. We adopted the common NEPS naming convention only for the first digit, generated variables and variables from competence tests.

The first digit indicates to which **primary respondent type** the variable refers, in case of the Organizational Reform Study in Thuringia this character can be "t" (target person), "p" (one parent of target person), "e" (educator in a specific course).

Additionally some information about the study structure and administration details are provided by the school coordinator - a selected teacher collaborating with the survey institute DPC. This information is sometimes essential to understand the data structure (e.g. which rotation or version of a test was administered for student x etc.) and therefore included into the specific data sets of the primary respondents. These variables are treated as generated variables and get an "x" as a second digit after the "t", "p" or "e".

Suffix (optional): Suffixes are separated from the previous characters by an underscore. There are three types of suffixes:

Suffixes for generated variables:

- Generated variables are indicated by the suffix *_g#* (*_g1*, *_g2*, etc.), *_ha*, *_v1*, and *_v2*. In most cases, the running number after *_g* is a simple enumerator. However, there are generated variables that assign meanings to these running numbers: occupational variables.
    - Occupational/prestige codes
        - g1: KldB 1988 (German Classification of Occupations 1988)
        - g2: KldB 2010 (German Classification of Occupations 2010)
        - g3: ISCO-88 (International Standard Classification of Occupations 1988)
        - g4: ISCO-08 (International Standard Classification of Occupations 2008)
        - g5: ISEI (International Socio-Economic Index of Occupational Status)
        - g6: SIOPS (Standard International Occupational Prestige Scale)
        - 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)
    - *_ha* indicates harmonized variables, which are generated from two variables which have nearly the same meaning in both years.
    - *_v1* and *_v2* indicate the original versions of the harmonized variables indicated with the *_ha* suffix.

- 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.

- Confidentiality suffix:

  This suffix pertains to all variables that were anonymized (see 1.4). The suffix indicates a variable's degree of anonymization. This suffix may either stand alone (e.g., country of birth: *p18am_R*) or be combined with other suffixes (e.g., coded nationality of the mother: *p19v_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

2.3 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 two parts and additional suffixes. The first part defines the test instrument (two/three characters, e.g. "ma" for Math), the second part defines the item number. There are two versions of item variables: scored items named *{varname}_c* and scored partial credit-items named *{varname}s_c*. Moreover, suffix *_sc{number}* is used for several scores and the meaning of the suffixed number is fixed as follows: 1=WLE (Weighted Maximum Likelihood estimates), 2=standard error of WLEs, 3=sum, 4=mean, 5=difference. If there are several versions e.g. of a sum score letters are appended additionally. For example, variable *mas2_sc1* represents the WLE score of the math test of students being tested. To give another example, variable *magcd541_c* is a scored version (values 0 or 1) of a math test item.

2.4 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 7). Table 3 provides an overview of missing codes you will encounter in the NEPS data.

Table 2 Overview of missing codes

Code	Missing
Item nonresponse	
–97	Refused
–98	Don't know
–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/–95	Implausible value removed (-52 assigned by data edition at NEPS Datacenter, -95 assigned by fiel work institute IAE-DPC)
–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).
  - For competence data there is a special missing code –94 that indicates that a test item has not been reached, because the target quits the test somewhere before this item.
  - Further missing codes (–20, …, –29) pertain to variable-specific nonresponse categories.
  - 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.
- *Edition missings* are defined in the process of data editing.
  - Implausible values are recoded into missing (–52) in the NEPS editioning process. Implausible values coded by a –95 missing have been removed already by the field work institute IAE-DPC.
  - 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 7 for further information on how to install and update the *nepsmiss* command.

3 Sampling and surveying procedures

3.1 Overview

Many German Länder are currently reforming the curriculum and the organization of the senior years of secondary school (*Gymnasiale Oberstufe*). In general, these changes target a stronger emphasis on general education and a restriction of the differentiation in the *Leistungskurs-Grundkurs-System* during the last two years in the *Gymnasiale Oberstufe*. The "NEPS – Organizational Reform Study in Thuringia" aims to study the possible effects of such a reform.

Two cross-sectional surveys were conducted for the graduation years 2010 (last year before the reform, NEPS study *A70*) and 2011 (first year after the reform, NEPS study *A71*) in Thuringia. The taret population of the study comprises all grade 12 students in 2010 und 2011 in Thuringia.

The students participated in a questionnaire, achievement tests (*Fachleistungstests*) in the fields of Mathematics, Physics, Biology, and English, and a test on cognitive abilities. In addition, relevant context persons were surveyed. That is, the students' parents and teachers for German, English, Math, Biology, Chemistry, and Physics were asked to complete a questionnaire. Field work was conducted by IEA DPC (IEA Data Processing and Research Center, Hamburg).

3.2 Sampling and response rates

The stratified sample[1] consists of all grade 12 students from 32 randomly selected grammar schools in Thuringia.

Table 3 Overview of samples and survey instrument participation

NEPS Study	**A70**	**A71**	
Study year (January)	2010	2011	pooled
Students (N-sampled initially)	1857	1365	3222
Students participated in questionnaire (PAPI)	1372	885	2257
Students participated in achievement tests	1374	886	2260
Parents participated in questionnaire (PAPI)	572	417	989
Teachers of selected courses participated in questionnaire (PAPI)[2]	407	300	707
Grading information available	1348	878	2226

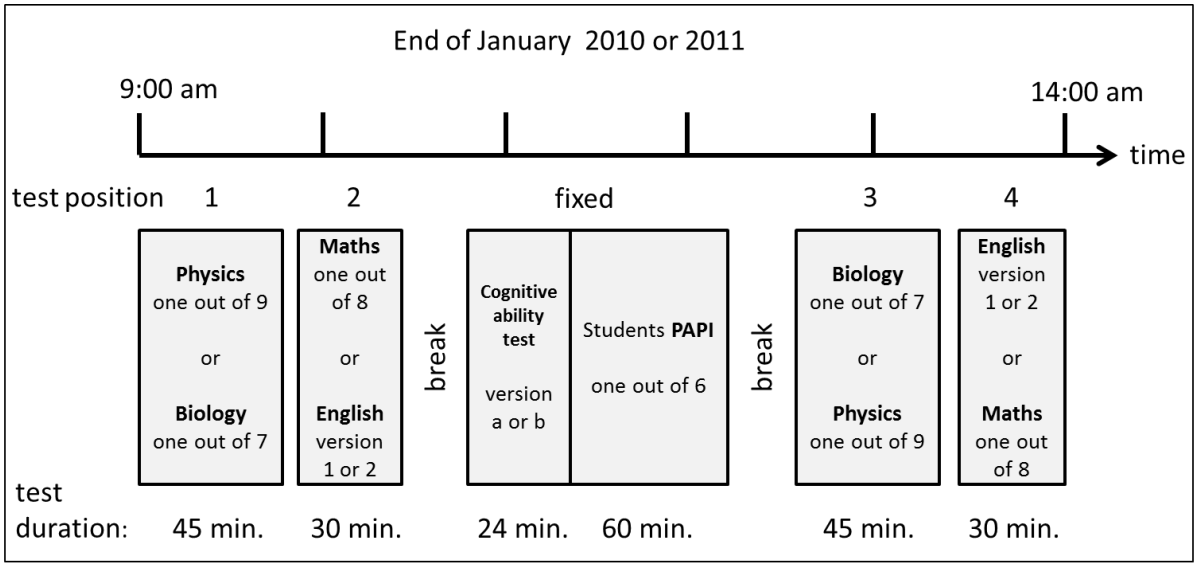
[1] For further details on the sampling and weighting procedure see technical report on weighting (Aßmann & Schönberger 2012)

[2] This implies the possibility that the same teacher provided information on different courses or in both years of the survey within one school. The teacher as an individual cannot be identified by design, but only his/her reference to a specific course. Hence, the number of teacher responses to questionnaires might be higher than the actual number of teachers providing this information.

3.3 Competence testing and students questionnaire

All achievement tests in the fields of Mathematics, Physics, Biology, and English, and the test on cognitive abilities as well as the students questionnaire (PAPI) where conducted in the schools within one day around end of January 2010 and 2011.

Figure 1 General procedure on students' competence/ability tests and surveying in school



In the first test session in the morning the students had to perform either in one out of nine different versions of a Physics test (see value *P1,...,P9* in the *id_phy* variable provided with the *xTargetCompetencies* data file) or in one out of seven different Biology test versions (value *B1,...,B7* in *id_bio*). This first block was intended to last 45 minutes. The second test session was dedicated to the domain of English (version *E1* or *E2* in id_eng) or Math (version *M1,...,M8* in *id_ma*) and was administered in 30 minutes.

After that – and a break – the students were asked to participate in the test on cognitive abilities (version *a* or *b* in *id_kft*; duration: about 24 minutes) and in the students questionnaire (version 89…41 in *tx80211*, duration about 60 mintes). After a second break, the third and fourth achievement tests were administered in the same manner as the first ones. Starting with either a Biology or Physics version and ending with Math or English forms again. To identify the relative position of a subject-specific test, separate order variables are provided in file *xTargetCompetencies* (*maorder,...,phyorder*). For a detailed description of the competence tests consult (Durchhardt (forthcoming) and Wagner (forthcoming)).

The students questionnaire consists of central socio-demographic questions (age, sex, country of origin, mother tongue etc.), questions on interests, aspirations, leisure activities, health, life satisfaction, the familial background from the students perspective, class climate as well as subject-related tuition and learning traits and finally on questions regarding reform aspects and consequences. The six versions differ mainly in the position of specific question blocks. Overall, the whole testing and survey procedure at the schools took around five hours.

3.4 Parents' questionnaire

To get further background information about the students and another perspective on the school and the reform in the senior years, the parents of the students were asked to fill out a PAPI questionnaire. Beside socio-demographic basics, this instrument focuses on reform-related opinions and ratings, aspirations and evaluations of their child and in the context of the reform.

3.5 Teachers' course questionnaire

All course teachers who are responsible for the subjects of participating students in German, English, Math, Biology, Chemistry, and Physics in the 12 grade were asked for *course specific* evaluations – also in respect to the reform and finally for some school- and person specific background information. Note however that the questionnaire (PAPI) is course specific and not a unique teacher instrument. Hence teachers may have provided information to more than one questionnaire, if they teach several subjects/courses or 12th graders in both years of the survey.

3.6 School grades and tracking information

School grades from all subjects and all four terms of the senior years of all 12th grade students of the participating schools as well as the performance levels of all subjects, the grading in the final exams and the final grade point average were collected. This was done retrospectively by the IAE-DPC at the end of the 12th grade in 2010 and 2011 via the school coordinators, the school principals and the data bases of the schools. The grades and results of those students not participating in the study had been send and processed in a completely anonymized and aggregated form and were used for the calculation of weights by the method group of the NEPS only. See technical report on weighing (Aßmann & Schönberger 2012) for details.

Additionally the students' sex, course type, course participation, instrument version or rotation respectively, and legal age was collected during the tracking process via the school coordinators (local teachers or principals responsible for organizing all activities within schools and classes that are necessary to realize the school survey). Similar tracking data on all course-teachers were also collected and send back alltogether to the DPC as annonymized lists.

4 Datafiles

As introduced above, the NEPS – Organizational Reform Study in Thuringia collects data of different types and from different sets of respondents: student data (paper questionnaire, competence and ability tests), parent data (take home paper questionnaire), course teacher data (paper questionnaire), tracking data from school coordinators, and, finally, students' grades provided by schools.

Except from the grading data and some tracking information (which are mainly integrated into the so called *Profile* dataset) all type/respondent data resemble into a separate dataset. In order to provide a most convenient data structure, the data from the two different cross sections in 2010 und 2011 are pooled in one file. The *Profile* dataset contains an indicator variable *wave* that identifies the pre/post-reform data, which can easily be merged to all other datasets. Remember, however, that this is not a panel wave indicator like it is in the datafiles of the starting cohorts of NEPS, since the Organizational Reform Study in Thuringia asked each student only once![3]

4.1 Pooled cross sectional target file: *xTarget*

The file *xTarget* contains all the data from the students' questionnaire as well as some information on the version of the administered instrument (*tx80211*), harmonized and original versions of some items (those with suffix *_ha, _v1, and _v2*) and coded educational and occupational aspirations and parents occupations (suffix *_g1,...,_g16*).

4.2 Pooled cross sectional competencies file: *xTargetCompetencies*

This file contains scored and scaled[4] data from the competence tests in Math, Physics, English and Biology, as well as the test data on general cognitive abilities. To facilitate the usage, instrument IDs (indicating rotation or version type) are included here from the tracking lists (*id_bio,...,id_kft*). Moreover the relative position of the four competence test during test day (see Figure 1) and the general participation indicator (tx_comp) and (tx_sfbkft) are integrated in the file as well.

4.3 Pooled cross sectional parent file: *xParent*

This file integrates the parents paper questionnaire responses in both cross sections and is enriched with coded educational and occupational scales (suffix *_g1,...,_g16*, see also Section 5.1).

[3] However the schools are sampled only once for both waves, which might lead to the situation that perhaps in some cases the same course teachers are asked in both waves – but for different courses in grade 12. Furthermore, if parents might have two or more children of only one year difference in age in the same school, also parents might have been asked twice – but for different target persons and partly for different topics. The design of the study does not allow the reidentification of teachers or parents in wave two, so there is no way to deal with this special type of clustering.

[4] WLE scores etc. for competencies are only availabel for maths so far. For English, Biology, and Physics these scores will be included in later releases.

4.4 Pooled cross sectional course file: *xCourse*

For convenience, responses to the different course teachers' questionnaires are integrated into one file for all types of courses (German, English, Math, Biology, Chemistry, and Physics). Unique course teachers' *IDs* (*ID_c*) are provided and separately available for each subject to facilitate merging. A separate subject indicator (subject) and requirement level (*tx_niveau*) originally collected via the course-teacher traking lists is also available here. See Figure 2 for an exemplary data snapshot.

Figure 2 Data example from *xCourse*

ID_c	subject	ID_cger	ID_cen	ID_cmat	ID_cphy	ID_cbio	ID_cch	tx_niveau
1007763	Englisch	.	1007763	.	.	.	.	erweiter
1007764	Physik	.	.	.	1007764	.	.	erweiter
1007765	Deutsch	1007765	.	.	.	.	.	Grundfac
1007766	Mathemat	.	.	1007766	.	.	.	erweiter
1007767	Biologie	.	.	.	.	1007767	.	erweiter

4.5 Linking and method file: *Profile*

To facilitate usage and enable a quick overview on the different types of data and cases available we provide a so called *Profile* dataset. This dataset contains the case number and IDs of all target persons (the students) who participated – at least in one instrument. It is therefore recommended also as a linking file for merges between different respondents. Additionally, the following central information is provided:

- *study* (NEPS-identifier of the study – A70 or A71).
- *ID_t* (unique student ifentifier).
- *ID_i* (the unique school identifier[5]).
- ID_c{ger|...|ch} for each subject/course teacher (e.g., *ID_cger, ID_cen*).
- wave (indicating the two separate cross sections for school year 2009/2010=1 and 2010/2011=2).
- Weights (nonresponse, design and total weights) plus standardized weights (suffix *_std*).
- Sex and being of legal age (*tx_sex* and *tx_vollj* (legal age, coded 2 if age < 18 and 1 if age >=18)) from the tracking information.
- Variables indicating whether data from a specific type of instrument, mode, or respondent is available, that is: achievement tests, students' questionnaire

[5] Unique only within wave!

combined with the test on general cognitive ability[6], parents' questionnaire, grading and course teachers' questionnaire (*tx_comp, tx_sfbkft, tx_efb, tx_grading, tx_ctger,…,tx_ctbio*[7]).

- Additionally, the grade point information ("Kurspunkte") from the school data bases for all subjects the student was enrolled (e.g., *ts24g1* provides grade points in Math course from the first half year in grade 11).

4.6 Clustering and merging within a multilevel data structure

The data structure resulting from this study has a medium multilevel complexity for school studies. Although there are many different possibilities to construct levels or clusters within the data, some emerge directly from the analytical, institutional and procedural perspective and should be of special interest. The following table gives a short overview on central perspectives and the main levels of interest.

Of course the same level might be of interest for more than one perspective.

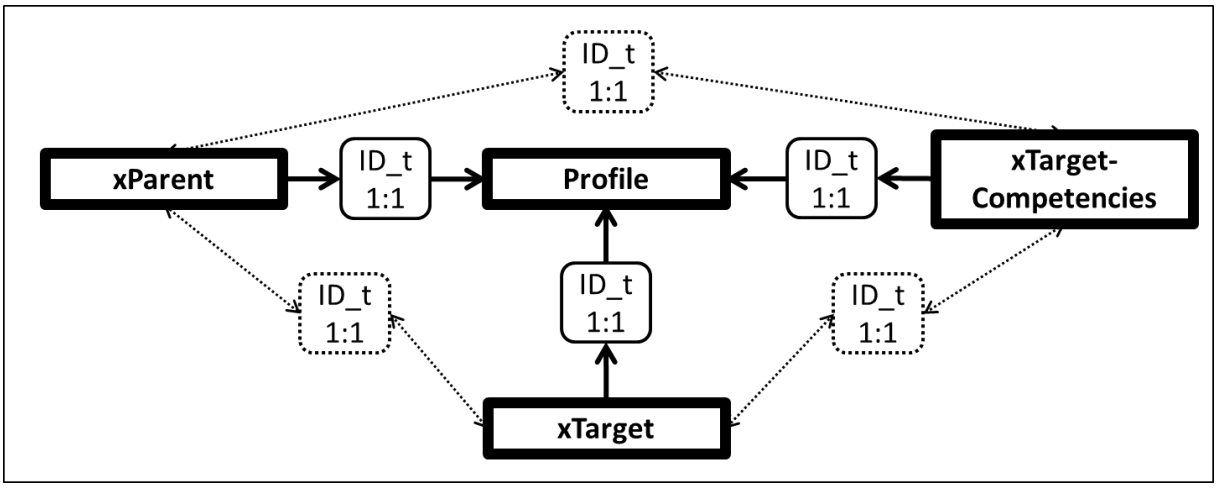
Table 4 Different general perspectives and possible levels/clusters within the data

Analytical	• Wave (*wave*) • Requirement levels (e.g., *ts11p*: German as "Leistungsfach") • Individual (*ID_t*)
Institutional	• School (*ID_i*) • Subject (e.g., Math) • Requirement level (e.g., ts11p: German as "Leistungsfach") • Specific course (e.g., *ID_cmat*) • Individual (*ID_t*)
Procedural	• Wave (*wave*) • School (*ID_i*)[8] • Rotation order/position (order at the day of survey; e.g., *maorder*) • Instrument version or rotation (e.g., *id_ma*)

[6] The test on general cognitive ability was administered as fix part of the students PAPI and hence there is only one indicator variable on participation for both instruments.
[7] A single indicator for the availability of course-teacher data would be misleading, since there are more than one (at maximum six) respondents (teachers) to this instrument for each student.
[8] There are also two different strata of schools (schools with focus on natural sciences vs. the rest). To accommodate to this sampling structure one can either use the provided design weights or create a stratum identifier on base of this weight (there are only two values). For further details see technical report on weighting (Aßmann & Schönberger 2012).

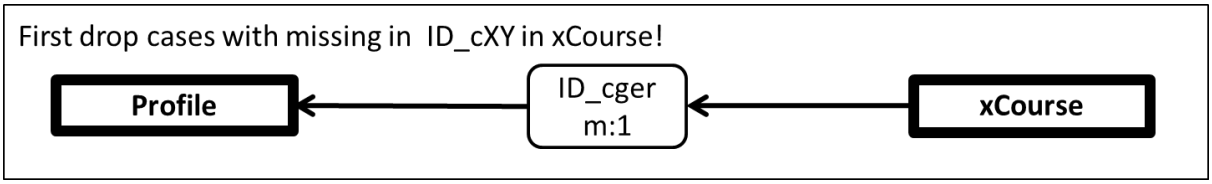
The most simple and probably most common type of merge is between the *Profile* data and the *xTarget*, *xParent,* or *xTargetCompetencies* files. Using ID_t as a unique identifier and a one to one merging command is enough to perform the merge. To keep all cases in this step, it is always recommended to merge against *Profile* first. (see section 6 for Stata examples).

Figure 3 Recommended merging relation for *Profile*, *xTarget, xParent*, and *xTargetCompetencies*



Linking the data from the *xCourse* file with the other datasets requires a different procedure. If course-teacher information is only needed for one specific subject (e.g. German), then a simple many to one merge is possible to the Profile data via the subject specific course teacher ID (e.g., *ID_cger*). After that the enriched *Profile* data can be merged with all the other data files with simple one to one merges as described above. However, it is necessary to drop the missings in the course teacher ID (e.g., *ID_cger*) in the *xCourse* data before the merging, since otherwise the ID would not be unique and a one to many merge would fail.

Figure 4 Merging data from one specific course-teacher (e.g. German)



If more than one course teacher information is needed at the students' level, again the merging strategy has to be modified. There are of many possible strategies to merge more or even all course-teacher data to the *Profile* and all other data files, but the following appears the most convenient to us. The description is made for a merge of data from all six course teachers, though the strategy remains the same for two to six.

- In a first step the *Profile* data has to be reshaped into a long format (student-course format). Each student data identified by ID_t represented by one data row in the original data has to be duplicated six times, since there are at maximum six course-teachers per student.

- Afterwards a many-to-one merge has to be done via ID_c – the teacher-course identifier.

- To get the data back in the original and preferred wide format (one row per student), a second reshape is necessary.

- In this step, all course teacher variables have to be renamed e.g. by adding a distinct suffix like "ger" indicating the type of course the data belongs to (in Stata this can be done within one single step, see Example 3).

- Finally, this integrated file can be merged with all the other files (*xTarget, xParent*, or *xTargetCompetencies*) via *ID_t* and a one to one merging procedure (see examples for the whole process in section 6).

Note

The NEPS invested a lot to ensure the integrity of these data. However, we strongly recommend you to examine the data critically when you work with this release. Furthermore, you should always consult the questionnaire/s to obtain a precise understanding of how the data have been collected.

5 Generated variables and weights

5.1 Coding

All string variables based on occupations, vocational training information and subjects of study of the respondents and their parents were coded. Table 4 presents an overview on these coded variables and the variables that are derived from them as well as the educational classifications (ISCED, CASMIN, years) which are particularly useful if you are interested in cross-national comparisons.

Table 5 Overview of coded variables

Classification	Included in	Description
KldB88	*p26m_g1*, *p26v_g1*, *p9_g1*, *t80a_g1*, *t80ba_g1*, *t80bb_g1*, *t80c_g1*	German Classification of Occupations 1988 (4-digit)
KldB2010	*p26m_g2*, *p26v_g2*, *p9_g2*, *t80a_g2*, *t80ba_g2*, *t80bb_g2*, *t80c_g2*	German Classification of Occupations 2010 (5-digit)
ISCO-88	*p26m_g3*, *p26v_g3*, *p9_g3*, *t80a_g3*, *t80ba_g3*, *t80bb_g3*, *t80c_g3*	International Standard Classification of Occupations 1988 (4-digit)
ISCO-08	*p26m_g4*, *p26v_g4*, *p9_g4*, *t80a_g4*, *t80ba_g4*, *t80bb_g4*, *t80c_g4*	International Standard Classification of Occupations 2008(4-digit)
BLK	*p26m_g9*, *p26v_g9*, *p9_g9*, *t80a_g9*, *t80ba_g9*, *t80bb_g9*, *t80c_g9*	Occupational classification by Blossfeld based on KldB92 (cf. Blossfeld 1985; Schimpl-Neimanns 2003)
ISEI-88	*p26m_g5*, *p26v_g5*, *p9_g5*, *t80a_g5*, *t80ba_g5*, *t80bb_g5*, *t80c_g5*	Metric scale to measure the socio-economic status of occupations based on ISCO-88 (cf. Ganzeboom et al. 1992; Ganzeboom 2010)
ISEI-08	*p26m_g14*, *p26v_g14*, *p9_g14*, *t80a_g14*, *t80ba_g14*, *t80bb_g14*, *t80c_g14*	Metric scale to measure the socio-economic status of occupations based on ISCO-08 (cf. Ganzeboom et al. 1992; Ganzeboom 2010)
SIOPS-88	*p26m_g6*, *p26v_g6*, *p9_g6*, *t80a_g6*, *t80ba_g6*, *t80bb_g6*, *t80c_g6*	Metric scale to measure prestige of occupations based on ISCO-88 (cf. Treiman 1977)
SIOPS-08	*p26m_g16*, *p26v_g16*, *p9_g16*, *t80a_g16*, *t80ba_g16*, *t80bb_g16*, *t80c_g16*	Metric scale to measure prestige of occupations based on ISCO-08
MPS	*p26m_g7*, *p26v_g7*, *p9_g7*, *t80a_g7* , *t80ba_g7*, *t80bb_g7*, *t80c_g7*	Magnitude prestige score of occupations (cf. Wegener 1985)
EGP	*p26m_g8*, *p26v_g8*	Class scheme which assigns occupations to classes (Erikson et al. 1079)
CAMSIS	*p26m_g15*, *p26v_g15*	Classification to measure social interaction and stratification (Prandy 2000)
CASMIN	*p20m_g2*,*p20v_g2*	Classification representing differentiated educational attainment and vocational training degrees

ISCED-97	*p20m_g1,p20v_g1*	Classification representing differentiated educational attainment and vocational training degrees (UNESCO 2006)
Years of education	*p20m_g3,p20v_g3*	Years of education based on the CASMIN classification

5.2 Weights

Generally three different kinds of weights are provided in the *Profile* dataset: design weights (*weight_design*), which can be used to correct for the stratified sampling, adjusted weights (*weight_adj*), which may be used to control for selective individual participation and finally a combination of both – total weights (weight_total). Furthermore the standardized versions (**_std*) of those three types of weights are also given for convenience. Detailed information on the construction of weights and how to use them can be found in the technical report on weighting (Aßmann & Schönberger 2012). For a more general discussion on the usage of sampling weights for model estimation see Rowher (2011).

6 Examples

This section gives some examples of how to merge different datasets. We provide you with the code to run the examples in Stata.

6.1 Example 1 – Merging data from *xParent* and *xTarget* via *Profile*

In the example shown below we merge data on the respondent's professional aspirations after graduation measured in terms of socio-economic status (e.g., ISEI 2008 level in *t80a_g5* in *xTarget*) and the father's actual ISEI 2008 level (*p26v_g14* in *xParent*) to the *Profile* data.

Code example 1 Merging information from *xParent and xTarget* data via *Profile*

```
*Merge specific information from xParent to xTarget via Profile

/*
Procedure
1. Open Profile
2. Merge variables from xParent to Profile with a 1:1-merge
3. Keep the relevant variables
4. Merge variables from xTarget to Profile with a 1:1-merge
*/

use "TH_Profile_D_1-0-0.dta", clear

merge 1:1 ID_t using "TH_xParent_D_1-0-0.dta ", keepusing (p26v_g14) nogen

keep ID_t p26v_g14

merge 1:1 ID_t using "TH_xTarget_D_1-0-0.dta", keepusing (t80a_g5) nogen

* recode missing values
nepsmiss *

* calculate a correlation
pwcorr t80a_g5 p26v_g14, sig
```

6.2 Example 2 – Merging *xTarget* with specific *xCourse* data

In the following example we merge the complete course-teachers information from German teachers to the *xTarget* and *xTargetCompetencies* data.

Code example 2 Merging information from xTarget with specific *xCourse* data

```
* Merge German course-teachers information from xCourse to
* xTarget and xTargetCompetencies via Profile

/*
Procedure
1. Open xCourse
2. Drop missing values in ID_cger and save the new file coursegerman
3. Open Profile
4. Merge variables from xCourse_ger to Profile with a m:1-merge
5. Merge with xTarget with a 1:1-merge and ID_t
6. Merge with xTargetCompetencies with a 1:1-merge and ID_t
*/

use "TH_xCourse_D_1-0-0.dta", clear

drop if missing(ID_cger)

tempfile coursegerman
save `coursegerman'

use "TH_Profile_D_1-0-0.dta", clear

merge m:1 ID_cger using `coursegerman', nogen

merge 1:1 ID_t using "TH_xTarget_D_1-0-0.dta ", nogen

merge 1:1 ID_t using "TH_xTargetCompetencies_D_1-0-0.dta ", nogen
```

6.3 Example 3 – Merging *xTarget* with *xCourse*

In the example shown below we merge all course teachers data *xCourse* to the *xTarget* and *xTargetCompetencies* data.

Code example 3 Merging information from xTarget with *xCourse*

```
*Merge all information from xCourse to xTarget and xTargetCompetencies via Profile

/*
Procedure
1. Open Profile
2. Reshape the data from wide to long using ID_t and ID_cger…ID_cch
3. Merge variables from xCourse to Profile with a m:1 and ID_c –merge
4. Drop unnecessary variables (ID_cger,...,ID_cch)
5. Reshape the data to wide format again
6. Merge with xTarget with a 1:1 -merge and ID_t
7. Merge with xCompetencies with a 1:1 -merge and ID_t
*/

use "TH_Profile_D_1-0-0.dta", clear

* reshape to student-course long format
reshape long ID_c@, i(ID_t) j(coursetype) string

merge m:1 ID_c using "TH_xCourse_D_1-0-0.dta", nogen

* drop redundant identifiers
drop ID_cger-ID_cch

reshape wide ID_c subject-e23_D, i(ID_t) j(coursetype) string

merge 1:1 ID_t using "TH_xTarget_D_1-0-0.dta", nogen

merge 1:1 ID_t using "TH_xTargetCompetencies_D_1-0-0.dta ", nogen
```

7 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
```

Furthermore, we have developed two 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)
```

nepsmiss: Recoding missing values

This program automatically recodes and labels all missing values into extended missing values (.a, .b, etc.). In this example, we run nepsmiss on the variable *t80a_g5*, decoding all negative values (-55, -97) into STATA's extended missings (.c, .a).

```
nepsmiss t80a_g5
```

ID_t	t80a_g5
5006300	-97
5006303	-55
5006304	-97
5006306	71
5006307	88

ID_t	t80a_g5
5006300	.a
5006303	.c
5006304	.a
5006306	71
5006307	88

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

infoquery: Display survey questions

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

```
infoquery t80a_g5
```

```
query result for variable t80a_g5:

t80a_g5[questiontext_de]:
Wenn es allein nach Ihren Wünschen ginge: Was würden Sie imAnschluss an den Schulabschluss
> (und gegebenenfalls Zivildienst,Wehrdienst, Soziales Jahr etc.) am liebsten machen?

t80a_g5[variablequestion_de]:
[nur Studium] Studienfach/Studienfächer:
```

8 Further information

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

https://www.neps-data.de/de-de/datenzentrum/forschungsdaten/zusatzstudiethüringen
(german)

https://www.neps-data.de/en-us/datacenter/researchdata/additionalstudythuringia
(english)

For further support, please contact the NEPS data center:

Web:

https://portal.neps-data.de/de-de/datenzentrum/kontaktdatenzentrum.aspx

E- Mail:

userservice.neps@uni-bamberg.de

Phone:

+49-(0)951-863-3511 (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.

References

Aßmann, C. & Schönberger, B. (2012). Weighting the Thuringia Samples of the National Educational Panel Study (NEPS). Technical Report. NEPS Research Data Papers, University of Bamberg.

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.

Blossfeld, H.-P. (1985). Bildungsexpansion und Berufschancen. Frankfurt: Campus.

Duchhardt, C. (forthcoming). Scoring and scaling of Math competencies in NEPS additional study "Organizational Reform Study in Thuringia". Technical Report, IPN Kiel.

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.

Ganzeboom, H. B. G. (2010). Questions and Answers about ISEI-08. Available from http://home.fsw.vu.nl/hbg.ganzeboom/isco08/qa-isei-08.htm.

Ganzeboom, H. B. G. , de Graaf, P. M., Treiman, D. J., de Leeuw, J. (1992). A standard international socio-economic index of occupational status. In: Social Science Research 21 (1992). S. 1 –56.

Koberg, T. (*2012*). NEPS Additional Study, Organizational Reform Study in Thuringia, (TH), SUF Version 1.0.0, Data Manual [Supplement]: Anonymization. NEPS Research Data Paper, University of Bamberg.

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 from: https://www.neps-data.de/Portals/0/Working%20Papers/WP_IV.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.

Wagner, W. (forthcoming). Scoring and scaling of competencies in NEPS additional study "Organizational Reform Study in Thuringia". Technical Report, Tübingen.

Wegener, B. (1985). Gibt es Sozialprestige? In: Zeitschrift für Soziologie 14(3). S. 209-235.