

Starting Cohort 4: 9th Grade (SC4)  
SUF Version 1.0.0  
Data Manual [Supplement]:  
Anonymisation  
*Tobias Koberg*



SPONSORED BY THE



Federal Ministry  
of Education  
and Research

Copyrighted Material

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

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

Principal Investigator: Prof. Dr. Hans-Günther Roßbach

Vice Managing Director: Prof. Dr. Sabine Weinert

Executive Director of Research: Dr. Jutta von Maurice

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

# **Starting Cohort 4 of the National Educational Panel Study: Anonymisation procedures and statistical disclosure control**

Technical Report

Tobias Koberg

National Educational Panel Study  
University of Bamberg

*v 1.0 October 26, 2012*

## Preamble

This documentation gives an exhaustive explanation of all disclosure risk minimisation techniques applied before dissemination of the Starting Cohort 4 (School and Vocational Training). For a quick reference what is done to the datasets in detail and on which level you will find your desired information, please skip forward to appendix A, where all affected variables are listed.

## Specifications

To ensure the best possible confidentiality protection of individuals and individual micro data, the National Educational Panel Study complies with strict international standards. Operationalise those, they have been abstracted to the following two criteria:

1. the disseminated data has been transferred to so called *de facto anonymous data*. Identifiable information is coarsened or cut off and kept securely to minimise the risk of statistical disclosure.
2. the use of data is strictly confidential and for statistical purposes only. The closed contract only grants access to members of the scientific community. This contract has a vast amount of legal stipulations, one of them being a large fine which applies for the realisation of re-identification on purpose. Therefore, the disseminated data is highly protected by law and allows a more flexible range of available data.

To pick up the latter, the NEPS has made a huge effort regarding legal regulations to offer as much analysis power of data as possible. This *paradigm of information esteem* reveals the fact that conducted measures of statistical disclosure control are few. Also, if there really was a need for modification, only non-perturbative methods were used.

## Onion-shaped model

The NEPS grants the user three different modes of data access: (1) ***OnSite***, which stands for the opportunity to use the secured infrastructure made available at the NEPS in Bamberg, (2) ***RemoteNEPS***, which is a progressive remote access technology providing a virtual desktop, and finally (3) ***Download***, indicating the possibility to fetch data via a secure web portal.

These given access modes have been originated to allow anonymisation routines for a subtle differentiation of information. The three resulting levels of anonymisation define as follows:

- data provided ***OnSite*** is generally not further anonymised. However, even those data has been rendered *de facto anonymous*, for no disclosure risk to persist. All information contained remains completely sane. Although users have to deal with limited possibilities of data access (i.e. supervised import and export of their results), they are free to work with all data available at the NEPS in a secure environment.

- access via *RemoteNEPS* is considered equivalent to *OnSite*, hence most of the data stays complete.
- as *Download* is assumed to be the most hazardous access mode<sup>1</sup>, some more anonymisation techniques are done to the dataset.

Obviously this approach results in three different versions of all involved datasets. To enable a consistent structure, these data files always contain the entire set of variables; it is their content which differs through the three levels.

As normally there is no need to resign aggregated variables in the higher levels (i.e. *OnSite* or *RemoteNEPS*), those are already defined as a surplus to the original variable in the *OnSite*-version. Stepping down to *RemoteNEPS* the content of related variables too sensitive for this level is overwritten with an exclusive missing code – an operation which we define as *purging*. Note that system missing values are not affected, allowing the user to differ between value existence and nonexistence. This still is a valuable additional information. Same applies to *Download*.

While there is no explicit documentation to this fact, it should remain clear that this procedure accumulates, i.e. purged content under *RemoteNEPS* is therefore neither included in *RemoteNEPS* nor in *Download*.

This *onion-shaped* model provides both ease of (1) use of different sensitivity models (e.g. preparing an analysis using the *Download* dataset and conducting it afterwards using the *OnSite*-data) and (2) documentation, for the subject of documentation is the most sensitive level (*OnSite*), with *RemoteNEPS* and *Download* levels being a subset of these data.

The fourth layer *master* depicted below contains every material which is needed during data processing by the NEPS, but is not meant for the scientific community to be usable.

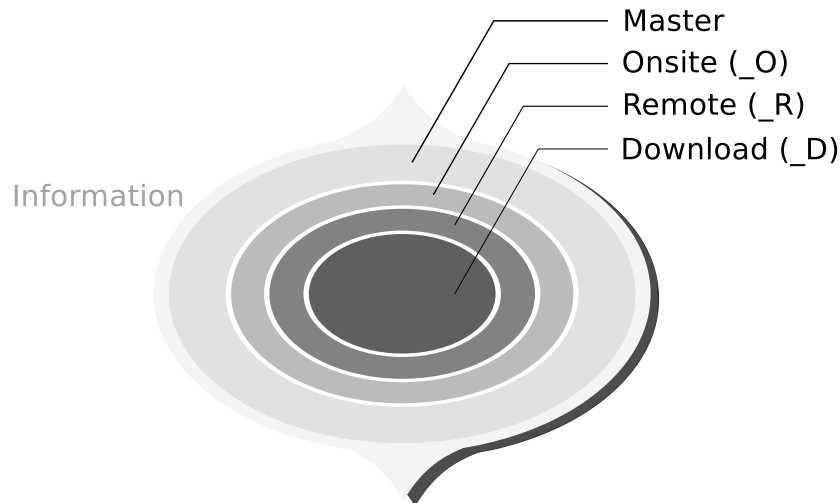


Figure 1: Onion-shaped model defining the different anonymisation levels

<sup>1</sup> ‘hazardous’ in terms of: the downloaded content is no longer under physical control of the NEPS

Technically, this model realizes in a single letter suffixed to dataset and variable names. All datasets available *OnSite* only are marked with an additional **\_O**, those available via *RemoteNEPS* with **\_R** and *Download* files with **\_D**. The same procedure applies when it comes to variable differentiation. A variable which is only available *OnSite* has been suffixed with **\_O**. In *RemoteNEPS*-access or *Download*, this variable is still present but purged. If there is an alternate version (mainly with coarsened content) for *RemoteNEPS* (suffix **\_R**) or *Download* (suffix **\_D**), those can be used. As said before, these are already integrated in the *OnSite* version.

## Conducted measures

Keeping the usability and the paradigm of information esteem in mind, only very few alterations are actually done to the dataset. These modifications always account for the fact that information may never be lost completely, but aggregated into coarse categories or variables. Please note that all information is still available somewhere and that only *RemoteNEPS* and (mainly) the *Download* version are constraint in this matter. In fact, roughly 120 variables are modified in some way – which is about six percent of the whole dataset volume.

Please refer to appendix A for a complete overview of all variables which fell victim to anonymisation.

The following gives an explanatory overview of all measures conducted.

**countries and languages** All information corresponding to (international) localisation, nationality or languages is only available en full *OnSite* or via *RemoteNEPS*. Variables comprised in the *Download* Scientific Use File (SUF) are aggregated into german and non-german.

**open ended strings** All string variables containing actual text are purged in the *RemoteNEPS* version. The information remains accessible *OnSite*. However, all text entries have been reviewed by staff to ensure that absolutely no re-identificational material is included.

**institutions** For starting cohort 2 to 4, special focus of anonymisation has been directed to protection of institutional data, i.e. information about kindergarten and schools, but also educators and teachers. This includes the complete datafile *xInstitution*, but also basic structural details about the kindergarten group or school class. Furthermore, personal information about educators and teachers is treated more securely. You will find detailed information about these subjects from *RemoteNEPS* onwards.

**regional Information** Regional information is not available for NEPS data which has been surveyed in school context. This regards places of birth as well as work, school or residence. Only an indicator for west germany and east germany (including Berlin) is available. Please be aware that we still do offer macro indicators *OnSite* (see below).

**number of employees** Considering self-employed persons, information about the number of salaried employees has been censored to prevent effortless identification of

large entrepreneurs. Therefore, related variables are top-coded at 20 employees. Again, this information is still present via *RemoteNEPS* and *OnSite*.

**macro indicators** Additional information including structural topography and macro-economic measures has been made available only *OnSite*, also called *RegioInfas (infas geodaten)*. Please refer to the separate documentation describing those datasets for further information.

Topic	<i>OnSite</i>	<i>RemoteNEPS</i>	<i>Download</i>
International <sup>1</sup>	full data	full data	collapsed
String variables	anonymised	n/a	n/a
Institutional	full data	full data <sup>2</sup>	n/a
Regional (national) <sup>3</sup>	collapsed	collapsed	collapsed
Number of employees	full data	full data	top coded
Macro indicators	accessible	n/a	n/a

<sup>1</sup> international geographical information (e.g., nation states, national languages)

<sup>2</sup> month of birth of educators/teachers and principals/headmasters is only available *OnSite*

<sup>3</sup> national localisation is coarsened to west/east germany

Table 1: Availability of sensitive data

For enquiries or further information not covered in this document please feel free to contact [userservice.neps@uni-bamberg.de](mailto:userservice.neps@uni-bamberg.de).

# A Anonymisation sheet

Instrument						
Name	Starting Cohort 9th grade (SC4), Version 1-0-0					
Stage	4					
Study type	Main survey					
Study number	A28_A56_A63_A46_A67_A83_A60_A86 / A47_A68_A84 / B34					
Dissemination	Scientific Use File					
Measures						
	data file	variable	label	On-site	RemotenePS	SUF-Download
<b>Countries &amp; Languages</b>						
ParentMethods	px80204		Interview: interview language (realized case)			purged
spSchool	p723090		Land of the school			purged
xTarget	i400000_g1		Country of birth			country aggregation
xTarget	i400070_g1		Mother: Country of birth			country aggregation
xTarget	i400090_g1		Father: Country of birth			country aggregation
xTarget	i400220_g1		Mother's mother: Country of birth			country aggregation
xTarget	i400240_g1		Mother's father: Country of birth			country aggregation
xTarget	i400260_g1		Father's mother: Country of birth			country aggregation
xTarget	i400280_g1		Father's father: Country of birth			country aggregation
xTarget	i40115a_g2		Nationality (response 1; coded)			country aggregation
xTarget	i40115a_g3		Nationality (response 2; coded)			country aggregation
xTarget	i40115a_g4		Nationality (response 3; coded)			country aggregation
xTarget	i42350x_g1		Definition culture of origin			country aggregation
xTarget	i41000a_g2		Mother tongue (1st alternative, ISO 639.2)			language aggregation
xTarget	i41000a_g3		Mother tongue (2nd alternative, ISO 639.2)			language aggregation
xTarget	i41000a_g4		Mother tongue (3rd alternative, ISO 639.2)			language aggregation
xTarget	i41000a_g5		Mother tongue (4th alternative, ISO 639.2)			language aggregation
xTarget	i41010a_g2		Mother: Mother tongue (1st alternative, ISO 639.2)			language aggregation
xTarget	i41010a_g3		Mother: Mother tongue (2nd alternative, ISO 639.2)			language aggregation
xTarget	i41010a_g4		Mother: Mother tongue (3rd alternative, ISO 639.2)			language aggregation
xTarget	i41010a_g5		Mother: Mother tongue (4th alternative, ISO 639.2)			language aggregation
xTarget	i41012a_g2		Father: Mother tongue (1st alternative, ISO 639.2)			language aggregation
xTarget	i41012a_g3		Father: Mother tongue (2nd alternative, ISO 639.2)			language aggregation
xTarget	i41012a_g4		Father: Mother tongue (3rd alternative, ISO 639.2)			language aggregation
xTarget	i41012a_g5		Father: Mother tongue (4th alternative, ISO 639.2)			language aggregation
xTarget	i410010_g2		Second language (1st alternative, ISO 639.2)			language aggregation
xTarget	i410010_g3		Second language (2nd alternative, ISO 639.2)			language aggregation
xTarget	i410010_g4		Second language (3rd alternative, ISO 639.2)			language aggregation
xTarget	i410010_g5		Second language (4th alternative, ISO 639.2)			language aggregation
xParent	p406010		Country of birth of target child			purged
xParent	p400010		Country of birth respondent			purged
xParent	p400090		Country of birth father respondent			country aggregation
xParent	p400070		Country of birth mother respondent			country aggregation
xParent	p401150		Nationality respondent not German			purged
xParent	p731804		Highest educational achievement abroad (country)			purged
xParent	p731823		Country of vocational qualification (additional response)			purged
xParent	p403010		Country of birth, partner abroad			purged
xParent	p403090		Country of partner's father			country aggregation
xParent	p403070		Country of birth of partner's mother			country aggregation
xParent	p404050		Other nationality partner			purged
xParent	p731854		Partner: Highest educational certificate abroad (country)			purged
xParent	p731873		Partner: Country of vocational qualification (additional response)			purged
xParent	p413000		First language/mother tongue of parent interviewed (list)			language aggregation
xParent	p413002		Further first language/ mother tongue of parent interviewed (list)			language aggregation
xParent	p414000		First language/mother tongue partner (list)			language aggregation
xParent	p414002		Further first language/mother tongue partner (list)			language aggregation
xParent	p410000		First language/mother tongue child (list)			language aggregation
xParent	p410002		Further first language/ mother tongue child (open)			language aggregation
xParent	p32601a		Position generator: Country - Nurse or male nurse			country aggregation
xParent	p32601b		Position generator: Country - Engineer			country aggregation
xParent	p32601c		Position generator: Country - Storekeeper or transport worker			country aggregation
xParent	p32601d		Position generator: Country - Social worker			country aggregation
xParent	p32601e		Position generator: Country - Sales clerk			country aggregation
xParent	p32601f		Position generator: Country - Police officer			country aggregation
xParent	p32601g		Position generator: Country - Physician			country aggregation
xParent	p32601h		Position generator: Country - Banker			country aggregation
xParent	p32601k		Position generator: Country - Car mechanic			country aggregation
xParent	p32601i		Position generator: Country - Legal practitioner			country aggregation
xParent	p32601m		Position generator: Country - Optician			country aggregation
xParent	p32601n		Position generator: Country - Translator			country aggregation
xParent	p32601o		Position generator: Country - Elementary, secondary modern or intermediate school teacher			country aggregation
xEducator	e41100a_g2		Mother tongue (response 1, ISO 639.2)			language aggregation
xEducator	e41100a_g3		Mother tongue (response 2, ISO 639.2)			language aggregation
xEducator	e41100a_g4		Mother tongue (response 3, ISO 639.2)			language aggregation
xEducator	e41100a_g5		Mother tongue (response 4, ISO 639.2)			language aggregation
<b>String variables (Note: all string variables have been approved for reidentificational material and anonymised where necessary)</b>						
spGap	ps29102		Other activity			purged
spSchool	p723090		School type (open)			purged
xInstitution	h229011		School: profile, other, text			purged
xInstitution	h229003		School: approach, pedagogic, text			purged
xInstitution	h229005		School: approach, promotion, text			purged
xInstitution	h229007		School: approach, integration, text			purged
xInstitution	h229009		School: approach, other, text			purged
xInstitution	h227011		School: Teaching staff, number of teachers each subject, other subjects, text 1			purged
xInstitution	h227012		School: Teaching staff, number of teachers each subject, other subjects, text 2			purged
xInstitution	h227013		School: Teaching staff, number of teachers each subject, other subjects, text 3			purged
xInstitution	h41700a		Other remedial teaching measures, other			purged
xInstitution	h41704a		Teacher training sessions migration, other			purged
xInstitution	h41702a		Parents education programs migration, other			purged
xInstitution	h222021		Quality assurance measures, other, text			purged
xParent	p731803		Highest educational achievement, respondent, type open			purged
xParent	p731814		Vocational qualification respondent (open)			purged
xParent	p731817		Type tertiary qualification, respondent (open)			purged
xParent	p731853		Highest educational qualification, partner, type (open)			purged
xParent	p731864		Vocational qualification partner (open)			purged
xParent	p731867		Type tertiary qualification partner (open)			purged
xParent	p262101		Parents: Tutoring, other subject (open)			purged
xParent	p269402		Wish: What vocation should the target child learn			purged
xTarget	td0026x		favorite subject			purged
xTarget	td27111		Courses outside school: Other courses (open)			purged
xTarget	td27270h		Student: Participation organization/group: Other, text			purged
xTarget	td26400		Student: Content of side job			purged
xTarget	td26350		Self-description sense of belonging			purged
<b>Size of class</b>						
xCourseClass	e227400_g1		Class: number of students total			aggregation
xCourseClass	e227400		class: amount of pupils, female			aggregation
xCourseClass	e227401		class: amount of pupils, male			aggregation
xCourseClass	e227402		class: amount of pupils, special pedagogic remedial teaching requirement			aggregation
xCourseClass	e451000		Number of students with a migration background in your class			aggregation
xCourseClass	e79201a		Students from lower social class (number)			aggregation
xCourseClass	e79201b		Students from middle social class (number)			aggregation
xCourseClass	e79201c		Students from higher social class (number)			aggregation
xCourseClass	e79202a		Students where at least one parent has graduated from college			aggregation
<b>Geographical region</b>						
xEducator	e537170_g1		location: passed the examination		east/west germany	
xEducator	e537110_g1		University teacher's course of study		east/west germany	
xEducator	e537030		Federal Land HZB		east/west germany	
xParent	p751001_g1		Place of Residence (RS West/East)		east/west germany	
spSchool	p723030_g1		Place of school (RS West/East)		east/west germany	
<b>Other</b>						
xEducator	e76212m		date of birth - month		purged	
xEducator	e76212y		date of birth - year			aggregation
xEducator	e229821		time in occupation - school			aggregation
xEducator	e229820		time in occupation - all in all			aggregation
xEducator	e536020		age career choice			topcoding
xEducator	e53710y		first enrollment teacher's course of study			aggregation
xEducator	e537150		Year of the examination			aggregation
xEducator	e53702y		Year HZB			aggregation
xParent	p731911		Number of employees respondent			topcoding
xParent	p731961		Number of employees, partner			topcoding
xCourseClass	e229400		class: fit-out: class room size			aggregation
xParent	p727001		Recommended secondary school/course of education (coarsened)	aggregation		
xParent	p727002		Further recommended secondary school/course of education (coarsened)	aggregation		
<b>Data files</b>						
xInstitution					not included	not included
xInstitutionRegionInfas					not included	not included
xTargetRegionInfas					not included	not included