NEPS
National Educational Panel Study

Research Data

Starting Cohort 2: Kindergarten (SC2) SUF-Version 1.0.0 Data Manual [Supplement]: Anonymisation

Tobias Koberg

SPONSORED BY THE

Federal Ministry
of Education
and Research

Starting Cohort 2 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 September 21, 2012

Preamble

This documentation gives an exhaustive explanation of all disclosure risk minimisation techniques applied before dissemination of the Starting Cohort 2 (From Kindergarten to elementary school). 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[1], 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.

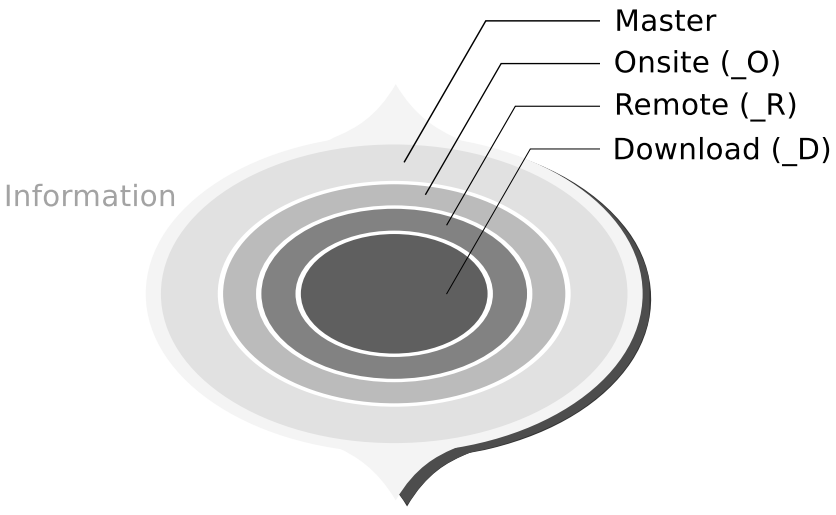

Figure 1: Onion-shaped model defining the different anonymisation levels

[1] '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 110 variables are modified in some way – which is about ten 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	*OnSite*	*RemoteNEPS*	*Download*
International[1]	full data	full data	collapsed
String variables	anonymised	n/a	n/a
Institutional	full data	full data[2]	n/a
Regional (national)[3]	collapsed	collapsed	collapsed
Number of employees	full data	full data	top coded
Macro indicators	accessible	n/a	n/a

[1] international geographical information (e.g., nation states, national languages)
[2] month of birth of educators/teachers and principals/headmasters is only available *OnSite*
[3] 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`.

A Anonymisation sheet

Instrument

Name	Starting Cohort Kindergarten (SC2), Version 1-0-0
Stage	2
Study type	Main survey
Study number	A12 / B11
Dissemination	Scientific Use File

Measures

	Datensatz	Variable	Label	On-site	RemoteNEPS	SUF-Download
Countries & Languages						
	ParentMethods	px80204	Interview: interview language (realized case)			Purged
	xParent	p406010	Country of birth of target child			Purged
	xParent	p407050	Nationality of the target child			Country aggregation
	xParent	p407060	Second nationality of the target child			Country aggregation
	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	p403010	Country of birth, partner abroad			Purged
	xParent	p401150	Nationality respondent not German			Purged
	xParent	p731804	Highest educational achievement, respondent, 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	Highest educational certificate abroad			Country aggregation
	xParent	p731873	Partner: Country of vocational qualification (additional response)			Purged
	xParent	p413000	First language/mother tongue interviewed parent (list)			Language aggregation
	xParent	p413002	Further language/mother tongue interviewed parent (list)			Language aggregation
	xParent	p414000	First language/mother tongue partner (list)			Language aggregation
	xParent	p414002	Further language/mother tongue partner (list)			Language aggregation
	xParent	p410000	First language/mother tongue child (list)			Language aggregation
	xParent	p410002	Further language/mother tongue child (list)			Language aggregation
	xParent	p412001	Interactive language household detailed (list)			Language aggregation
	xParent	p731823	Country of vocational qualification (additional response)			Purged
	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
String variables ***(Note: all string variables have been approved for reidentificational material and anonymised where necessary)***						
	xEducator	e217406	Number of children with diagnosed disorders, other developmental disorders, text		Purged	
	xEducator	e21980b	Nursery school group leaqder: Professional qualification, other qualification, text		Purged	
	xEducator	e219817	Group leader: Extent of work in terms of: Other, text		Purged	
	xEducator	e212819	Advanced training, other, text		Purged	
	xInstitution	h217006	Kindergarten: Number of children with other diagnosed developmental disorders, text		Purged	
	xInstitution	h21920a	Kindergarten: Diagnostic offers, other, text		Purged	
	xInstitution	h21920b	Kindergarten: Therapeutic offers, other, text		Purged	
	xInstitution	h21920c	Kindergarten: Offers for parents, other, text		Purged	
	xInstitution	h212013	Participation in quality development measure, other measure, text		Purged	
	xInstitution	h21905t	Kindergarten: Problems with surroundings, other, text		Purged	
	xInstitution	h418012	Other individual promotion open		Purged	
	xInstitution	h418052	Other type of small-group promotion open		Purged	
	xInstitution	h418092	Other whole-group promotion - open		Purged	
	xInstitution	hb1006t	Execution of the speech promotion measure persons with other qualification, text		Purged	
	xInstitution	h40184a	Measures for parents with an immigration background open 1		Purged	
	xInstitution	h40185a	Measures intercultural competence open for educators (open)		Purged	
	xInstitution	h21932t	Staff, other qualification, text		Purged	
	xInstitution	h212307	Kindergarten: Frequency of honorary tasks, other, text		Purged	
	xInstitution	h219817	Use of actual weekly working hours, other, text		Purged	
	xInstitution	h212819	Advanced training, 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	
Size of class						
	xEducator	e219410	Nursery school core group: Room use: Number of rooms			TopCoding
	xEducator	e219411	Nursery school core group: Room use: Size of rooms			Aggregation
	xEducator	e217401	Nursery school core group: Girls registered			Purged
	xEducator	e217402	Nursery school core group: Boys registered			Purged
	xEducator	e217403	Core group: Number of children with diagnosed disorders, speech			Purged
	xEducator	e217404	Core group: Number of children with with diagnosed disorders, behaviour			Purged
	xEducator	e217406_g1	Core group: Children with other development disorders - specified			Purged
	xEducator	e217405	Number of children with diagnosed disorders, other developmental disorders			Purged
	xEducator	e217412	Core group: Year of birth 2009 and later; number of children altogether			Purged
	xEducator	e217422	Core group: Year of birth 2009 and later; number of children, attend up to 5 hours			Purged
	xEducator	e217432	Core group: year of birth 2009 and later; number of children, attend 5-7 hours			Purged
	xEducator	e217442	Core group: year of birth 2009 and later; number of children, attend more than 7 hours			Purged
	xEducator	e45110f	Core group: Year of birth 2009 and later; number of children, migration background			Purged
	xEducator	e217452	Core group: Year of birth 2009 and later; number of children, disability			Purged
	xEducator	e217413	Core group: Year of birth 2008; number of children, altogether			Purged
	xEducator	e217423	Core group: Year of birth 2008; number of children, attend up to 5 hours			Purged
	xEducator	e217433	Core group: Year of birth 2008; number of children, attend 5-7 hours			Purged
	xEducator	e217443	Core group: Year of birth 2008; number of children, attend more than 7 hours			Purged
	xEducator	e45110e	Core group: Year of birth 2008; number of children, migration background			Purged
	xEducator	e217453	Core group: Year of birth 2008; number of children, disability			Purged
	xEducator	e217414	Core group: Year of birth 2007; number of childre, altogether			Purged
	xEducator	e217424	Core group: Year of birth 2007; number of children, attend up to 5 hours			Purged
	xEducator	e217434	Core group: Year of birth 2007; number of children, attend 5-7 hours			Purged
	xEducator	e217444	Core group: Year of birth 2007; number of children, attend more than 7 hours			Purged
	xEducator	e45110d	Core group: Year of birth 2007; number of children, migration background			Purged
	xEducator	e217454	Core group: Year of birth 2007, number of children, disability			Purged
	xEducator	e217415	Core group: Year of birth 2006; number of children, altogether			Purged
	xEducator	e217425	Core group: Year of birth 2006; number of children, attend up to 5 hours			Purged
	xEducator	e217435	Core group: Year of birth 2006; number of children, attend 5-7 hours			Purged
	xEducator	e217445	Core group: Year of birth 2006; number of children, attend more than 7 hours			Purged
	xEducator	e45110c	Core group: Year of birth 2006; number of children, migration background			Purged
	xEducator	e217455	Core group: Year of birth 2006; number of children, disability			Purged
	xEducator	e217416	Core group: Year of birth 2005; number of children, altogether			Purged
	xEducator	e217426	Core group: Year of birth 2005; number of children, attend up to 5 hours			Purged
	xEducator	e217436	Core group: Year of birth 2005; number of children, attend 5-7 hours			Purged
	xEducator	e217446	Core group: Year of birth 2005; number of children, attend more than 7 hours			Purged
	xEducator	e45110b	Core group: Year of birth 2005; number of children, migration background			Purged
	xEducator	e217456	Core group: Year of birth 2005; number of children, disability			Purged
	xEducator	e217417	Core group: Year of birth 2004 and earlier; number of children, altogether			Purged
	xEducator	e217427	Core group: Year of birth 2004 and earlier; number of children, attend up to 5 hours			Purged
	xEducator	e217437	Core group: Year of birth 2004 and earlier; number of childre, attend 5-7 hours			Purged
	xEducator	e217447	Core group: Year of birth 2004 and earlier; number of children, attend more than 7 hours			Purged
	xEducator	e45110a	Core group: Year of birth 2004 and earlier; number of children, migration background			Purged
	xEducator	e217457	Core group: Year of birth 2004 and earlier; number of children, disability			Purged
	xEducator	e401200	Number of children with another interactive language			Purged
	xEducator	eb1000a	Language skills above average age - all children			Purged
	xEducator	eb1000b	Language skills above average age - thereof children with a migration background			Purged
	xEducator	eb1000c	Language skills average - all children			Purged
	xEducator	eb1000d	Language skills average - thereof children with a migration background*			Purged
	xEducator	eb1000e	Language skills below average - all children			Purged
	xEducator	eb1000f	Language skills below average - thereof children with a migration background*			Purged
	xEducator	e401200	Number of children with another interactive language			Purged
Other						
	xEducator	e76112m	Month of birth		Purged	
	xEducator	e76112y	Year of birth			Aggregation
	xEducator	e219800	Nursery school group leader: Professional qualification			Aggregation
	xEducator	e219820	Group leader: Duration of professional activity: Previous facilities			Aggregation
	xEducator	e219821	Group leader: Duration of professional activity: Current facility			Aggregation
	xInstitution	h76612m	Month of birth		Purged	
	xParent	p731911	Number of employees respondent			TopCoding
	xParent	p731961	Number of employees, partner			TopCoding
	xParent	p751001_g1	Place of Residence (RS West/East)	East/West Germany		
	xParent	pb11610_g1	Place of elementary school (RS West/East)	East/West Germany		
Data files						
	xInstitution					not included
	xInstitutionRegioInfas				not included	not included
	xTargetRegioInfas				not included	not included