NEPS WORKING PAPERS

Tobias Koberg and Katharina Stark

MEASURING INFORMATION REDUCTION CAUSED BY ANONYMIZATION METHODS IN NEPS SCIENTIFIC USE FILES

NEPS Working Paper No. 65
Bamberg, January 2016

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg − Germany − contact@lifbi.de

Measuring Information Reduction caused by Anonymization Methods in NEPS Scientific Use Files

Tobias Koberg, Katharina Stark
Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

tobias.koberg@lifbi.de

Bibliographic data:

Koberg & Stark (2016). *Measuring Information Reduction caused by Anonymization Methods in NEPS Scientific Use Files* (NEPS Working Paper No. 65). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Measuring Information Reduction caused by Anonymization Methods in NEPS Scientific Use Files

Abstract

The National Educational Panel Study (NEPS) is conducted as a very large panel study, collecting data from six Starting Cohorts in different age ranges. Primarily, NEPS is designed as an infrastructural service, providing the collected information to researchers as Scientific Use Files (SUF). Those SUFs are disseminated via three different access modes: Researchers may work with the data *Onsite* at our facility, they may use our remote access technology *RemoteNEPS*, or *Download* the data from our website to their local workstation. This strategy is used to protect more sensitive information by more secure access ways, which means every access way offers a specific SUF version, each containing a different amount of information. This is done by modifying the data later provided for Download and Remote usage, reducing the information contained to a more anonymous level (e.g., topcoding some variables), and thus being more appropriate for this level. In this paper, we try to measure those differences, that is, to determine a measurement of information difference, by quantifying the information loss when comparing the data. We do this following three approaches: (1) counting the amount of variables affected by anonymization, (2) evaluating the methods applied by an heuristic approach, and (3) measuring the difference of the empirical data. It turns out that by referring to the Onsite SUF versions as 100% (i.e., the full information is accessible here), on average between 74% and 87% of information is preserved in the Download, and more than 97% in the Remote versions.

Keywords

anonymization, information loss, Scientific Use File, data comparison

1. Introduction

The National Educational Panel Study (NEPS) surveys a vast amount of longitudinal data about individuals, context persons, and institutions, and has been established to provide the collected data to the scientific community. This mandate produces two (partially contrary) challenges: At first, individual (or personal) data must always be handled most considerately. Careful practice has to be applied when such data is passed to others, as any malicious handling may result in breaking the law or, at least, displease affected respondents, which may lead them to abandon the participation. This tends to support the idea of disseminating data very sparingly and with uttermost harsh anonymization. On the other hand, this project is an infrastructural service for researchers of heterogeneous disciplines. For NEPS, neither their research goal, nor the information they are interested in is easily foreseeable. Hence, if not all information is initially disseminated, it might prevent certain research projects of being conducted. NEPS tries to counter those two challenges by disseminating more *sensible data* (i.e., data which might reidentify individuals more likely than others) by more secure access ways. By this approach, both data security as well as data accessibility are satisfied.

1.1 Concept of anonymization

To ensure an effectual protection of the data, NEPS has established a so called *portfolio approach*. This approach interlocks five different security mechanisms to guarantee the best possible protection for the data:

- institutional (data is only disseminated to researchers)
- legal (data users have to sign a data use agreement)
- informational (users are sensibilized to data protection and potential misuses)
- technical (more sensible data is only available under access restrictions)
- and statistical (modifying data values and render them more anonymous)

While all of those approaches are more or less equally important, we only focus on the latter two of them in the setting of this paper: the technical and the statistical approach. Although the other approaches too regularize access to NEPS data, they do not immediately affect the amount of information. They can therefore be neglected. Nonetheless, it is crucial to note, that, without the other three approaches, the elaborated results out of this work would underestimate the need for and amount of anonymization. Those results are only meaningful while keeping in mind the strict settings of data dissemination. More about this general data protection concept can be found in Koberg, 2016.

By *technical approach*, we refer to the three access modes NEPS data can be obtained. The first access mode, Onsite usage, only allows users to work with the data at our location. As the data does not leave our custody, access can be supervised and controlled. In this highly regulated setting, we feel safe to allow access to more sensitive data. Yet, of course, all criteria are hold to preserve anonymity of persons, households and institutions.

The second access mode is our remote access solution *RemoteNEPS*. Now, the data does not physically leave our safeguarding, that is, it still remains inside our system. However, handling and further processing of the data can not be fully controlled, as the view on the data is transmitted to the users workstation. It is somewhat a hybrid solution, combining the benefits of a secured surrounding with the comfort of anywhere access. Hence, more information than in the Download version is available, but still not as much as Onsite.

Via the third access mode, data users may download the data from our website to their local workstation. As data leaves the secured area now (i.e., data files are copied out of our systems), the missing physical protection has to be balanced by additional anonymization. This is established by modifying the data to render them more anonymous, for example, clustering and topcoding or completely removing some information. This modification is the *statistical approach*.

In the data versions used for this paper (see table 1), only a handful of non-perturbative[1], global[2] modification methods were used in the course of the statistical approach. The following gives a short, yet complete overview of all applied methods:

aggregating values, so they are represented on a more coarse level. In literature, this is often referred to as *global recoding* (e.g., see Hundepool et al., 2012). For example, country of birth is recoded to Germany/Abroad, place of residence to East/West Germany, and so forth.

topcoding is a special case of aggregation, where only the end of the scale is affected. This could be the number or size of classrooms (1, 2, 3, 4 and more), number of employees (1,2,..., more than 50), and so forth.

bottomcoding is the same procedure applied to the lower end of the scale (e.g., year of birth before 1950).

aggregating and top-/bottomcoding in combination also occurs, for example, year of birth is aggregated to decades with an open highest category, size of class (i.e., amount of students in class) in some few categories (under 20, 20-25, over 25), and so forth.

percentualizing values, and removing total values. So, for example, instead of offering the total amount of girls in a class, this number is divided by the total amount of students (of both genders), generating the proportion of this.

removing information too sensible for this access mode. As an example, in NEPS data, regional data is only available Remote or Onsite. It is removed in the Download version.

Summing this up, for each of those three access modes, one version of a Scientific Use File (SUF) is generated, containing the amount of information appropriate for this access. At Onsite level, most information can be accessed, which also means, least data modifying anonymization has been conducted. At Download level, data has been anonymized most, resulting in less available information. The version available in RemoteNEPS contains more information than the Download version, but less than the Onsite version. While constructing those three different versions, a process referred to as *purging* was utilized. Rather than substituting sensible variables by their modified (more anonymous) counterpart in the Remote or Download version, both the original as well as the modified variable is kept in all three SUF versions. Valid values in the original variable are then overwritten with a specific missing code – the variable is purged. By this method, all variables can be accessed in all versions, but their content, the information within, stays secure if necessary.

1.2 Scientific Use Files

NEPS consists of six different panel cohorts (or Starting Cohorts, short SC) and two additional cross-sectional studies. Starting Cohort 1 (SC1) starts collecting data from infants in their early

[1] perturbative methods alter data values by overlying them with additional noise. See for example Oganian, 2011.

[2] while in *local* methods (e.g., *local suppression*, Hundepool et al., 2012), values of individuals are altered, *global* methods recode the whole variable for all individuals.

childhood (first year). In Starting Cohort 2 (SC2), the sample consists of 4-year-olds, visiting Kindergarten, while in Starting Cohort 3 (SC3) and 4 (SC4), students in secondary school are sampled (5th grade and 9th grade). Starting Cohort 5 (SC5) targets first-year university students and Starting Cohort 6 (SC6) focuses on adults and their life course. Additionally, two studies concerning school reforms were conducted in Thuringia (TH) and Baden-Wuerttemberg (BW). Although those are cross-sectional, these studies contain multiple waves, observing student cohorts before and after the reforms.

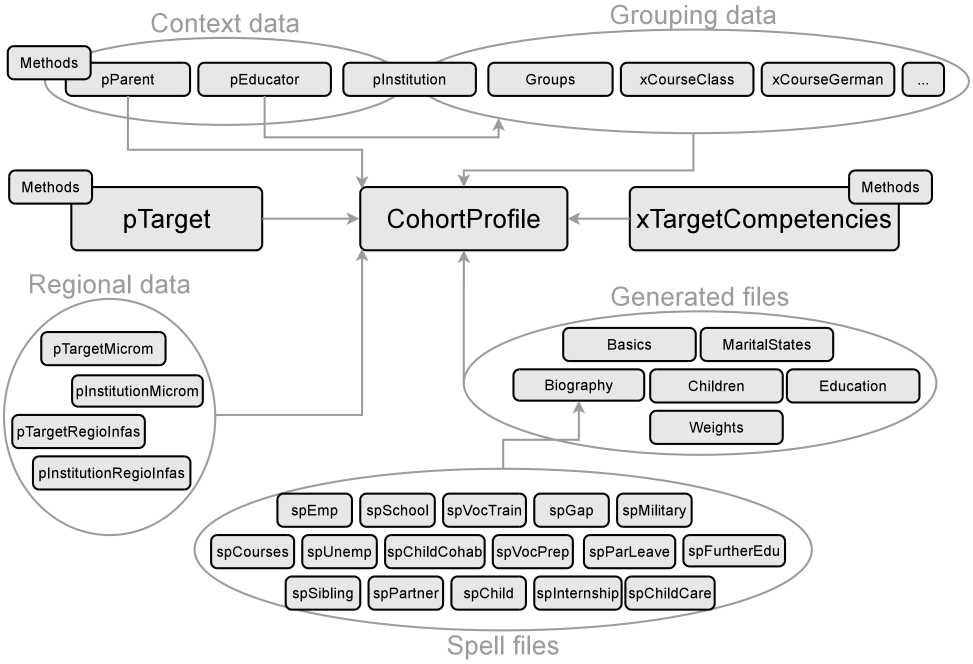

Figure 1. Generic representation of data sets included in NEPS Scientific Use Files

For every of this six plus two cohorts/studies, one SUF has been generated. A SUF is not a single file, as the name suggests, but rather a collection of datasets, or physical files, clustering information useful and easily accessible. Each data file then holds multiple variables of (more or less) related content. The total amount of real data files ranges from 4 files (in BW) up to 34 (in SC4). It has been tried to apply the same logic of file construction in all cohorts, so the structure of SUFs becomes mostly equivalent in all cohorts. However, as the surveys are very heterogeneous (both in content and in survey methods), so are the resulting SUFs. Figure 1 gives an overview of potentially enclosed files, although this is a generic representation (you may not find all of those files in every SUF, and the variables inside might not be same). Refer to the corresponding data manual of the survey (e.g., Skopek, Pink, and Bela, 2013 for Starting Cohort 4) for more about study background, sample, data files, their logic, and content information.

As NEPS is a panel study, respondents are surveyed in regular intervals (usually once a year). Rather than releasing a separate SUF for this newly collected information, the existing SUF is appended with the new wave and an update is then released. Therefore, still there is only one SUF for one Cohort, but it now contains both the information of previous and new waves. The increment of information contained (i.e., the additional wave) is indicated by a version number. Of course, as successive waves may collect completely new or different information, additional or completely new anonymization methods may become necessary. The analysis conducted in this working paper is based on the currently available SUFs. See Table 1 for an overview and volume of the underlying data.

In the remainder of this paper, we try to elaborate a way to quantify the information difference

Table 1
Analyzed SUFs of the Starting Cohorts, their version and some key values

	version	No. of waves	targets	files	variables
SC1	doi:10.5157/NEPS:SC1:2.0.0	2	3 481	10	2 168
SC2	doi:10.5157/NEPS:SC2:3.0.0	3	9 337	17	2 727
SC3	doi:10.5157/NEPS:SC3:3.1.0	3	8 317	18	3 743
SC4	doi:10.5157/NEPS:SC4:4.0.0	4	16 425	34	4 350
SC5	doi:10.5157/NEPS:SC5:4.0.0	4	17 913	26	2 573
SC6	doi:10.5157/NEPS:SC6:5.1.0	5	17 140	31	2 545
BW	doi:10.5157/NEPS:BW:3.0.0	3	5 210	4	1 520
TH	doi:10.5157/NEPS:TH:2.0.0	2	2 260	5	1 807

(i.e., reduction) between the SUFs of all cohorts in the three versions Onsite, Remote, and Download.

2. Quantifying anonymization amount

To do this, we first focus on the differences between the Onsite version (no modification) and the Download version (information removed due to modification) inside one Starting Cohort. For convenience, in the following we denote the information contained in the Onsite version $\mathcal{I}_O$, and the information contained in the Download version $\mathcal{I}_D$. In both versions, the same amount of variables is contained (see the concept of purging in section 1.1). We denote their amount by V, and refer to one variable as v_i, $i = 1, \ldots, V$. For our purpose, it is sufficient to assume that $\mathcal{I}_O$ is the complete information[3]. $\mathcal{I}_D$ has been target of some data reducing methods, so it contains only a subset of information, $\mathcal{I}_D \subset \mathcal{I}_O$. Our goal is to estimate the proportion $\mathcal{I} := |\mathcal{I}_D| \;/\; |\mathcal{I}_O|$, which measures the amount of information transferred.

2.1 Proportional approach

The first idea is to have a look at the number of variables affected by anonymization. It turns out that most variables in the data files are actually passed through unaltered. The percentage of affected variables varies across Starting Cohorts, but has its mean at 25% (see table 2). So for a rough estimate, one could take this proportion as a measure for $\mathcal{I}$:

$$\hat{\mathcal{I}}_P := 1 - \frac{\{v_i | v_i \text{ has been subject to anonymization}\}}{V}$$

This seems not very accurate, however, because simply being modified does not yet mean to lose all information. We therefore aim for a more elaborate way of quantification, by assigning a weight factor w_i to each variable v_i. To compute this weight, we try two different approaches.

2.2 Heuristic approach

In the heuristic approach, we do not consider the actual (surveyed) data of our respondents. We infer the amount of information loss solely out of the methods applied. For a variable v_i, the

[3] actually, this is not true, because there still is some information withheld, for example, auxiliary variables, which are not passed to the end user; this is, however, irrelevant for this comparison.

factor $w_i \in [0,1]$ shall represent the amount of information still contained in $\mathcal{I}_D$ of this variable. Once we have computed w_i for every variable, we can easily infer our aim by their mean:

$$\hat{\mathcal{I}}_H := \frac{1}{V}\sum_{i=1}^{V} w_i$$

In total, we have identified the following anonymization procedures, and therefore, the following options to compute our weights:

full loss/no loss If a variable v_i is completely removed during anonymization, that is, there is null information preserved, we set $w_i = 0$; on the other side, when the variable is not affected at all, full information is preserved, so we set $w_i = 1$.

aggregation A common strategy to anonymize information in one variable is to transfer values to a more general domain. Usually, this is done by grouping multiple similar categories to one global category, which then represents all associated values (see section 1.1 for some examples).

To compute w_i for this procedure, we let K_i be the total number of categories in the variable v_i – as it can be found in $\mathcal{I}_O$. Now, when observing $\mathcal{I}_D$, the domain set and therefore the information in some variables has been aggregated into global categories. Let's say the variable v_i has now G_i global categories g ($g = 1, \ldots, G_i$), each comprising k_g (sub)categories[4]. We then compute w_i as $w_i := \frac{G_i}{K_i}$. This computation is applied to all variables where K_i is known and bounded.

top-/bottomcoding Sometimes, a variable is aggregated by topcoding their highest values, for example, something like *»1,2,...,x and above«*. Now, we do not know K_i (as *»and above«* is not bounded), and therefore can not apply the previous calculation. So we came up with an approximation:
Generally, topcoding is conducted to summarize the topmost outliers, that is, only the highest category G_i is affected. To reproduce this, we suppose (!) that the values/categories still preserved in $\mathcal{I}_D$ divide the value set by more or less equal fractions, assuming that every category holds the same amount of information. We therefore define our weight as 1, reduced by the information aggregated in the last (top-coded) category: $w_i = 1 - 1/G_i$. We use the same logic for bottomcoding.

Sometimes, both aggregation and topcoding was conducted. Here, we assume that the total bandwidth is prolonged by the mean fraction size of all other categories, that is we estimate k_{G_i} by $\hat{k}_{G_i} := \frac{1}{G_i - 1}\sum_{g=1}^{G_i - 1} k_g$. With this, all variables are known to compute w_i using the formula for aggregation.

percentages In some cases, total values have been replaced by their percentage. In the computation of those, three values appear: N, the count of the total population; n, the count of a subpopulation; and n/N, the percentage of this. You easily can derive each of those numbers by the other two, so one of those is redundant. During the anonymization of $\mathcal{I}_O$ to $\mathcal{I}_D$, the variable v_{i_N} containing N is removed (as there is no reference value to compute a percentage), and the variable v_{i_n} containing n is being replaced by n/N. Because of the redundancy and already accounting for the erasure of N by setting $w_{i_N} = 0$, we do not determine any information loss in v_{i_n}, so $w_{i_n} = 1$.

[4] as transferring one category from $\mathcal{I}_O$ to $\mathcal{I}_D$ unaltered is equivalent to having $k_g = 1$, we do not have to separately account for this case.

To summarize this, we have:

$$w_i := \begin{cases} 1 & \text{if } \mathcal{I}_O(v_i) = \mathcal{I}_D(v_i) \text{ or percentages} \\ \frac{G_i}{K_i} & \text{if aggregation} \\ \frac{G_i}{K_i^*} \; with \; K_i^* := \hat{k}_{G_i} + \sum\limits_{g=1}^{G_i - 1} k_g & \text{if aggregation and topcoding} \\ 1 - \frac{1}{G_i} & \text{if topcoding} \\ 0 & \text{if } \mathcal{I}_D(v_i) = \emptyset \end{cases}$$

2.3 Empirical approach

Instead of considering a potential anonymization amount, one could also fall back to the actual data. To motivate this, reflect the considerations above. We would estimate the amount of preserved information in an aggregation of the variable *country of birth* (in total $K = 222$ countries) to *»1=Germany/2=Abroad«* by $w_i = \frac{2}{222}$. Yet, we have to account that NEPS is a national survey, and therefore a high percentage of respondents (partially over 90%) was born in Germany. As this information is completely preserved in $\mathcal{I}_D$, such a small weight seems odd. One would rather expect w_i being close to 1.

To utilize our data, we determine an estimator $\hat{w}_i$ of our weight by comparing the empirical probability distribution of the same variable v_i between $\mathcal{I}_O$ and $\mathcal{I}_D$. We call them p_{i_O} and p_{i_D}. We make the following assumptions about p_i:

a) the distributions are discrete. Almost all variables in NEPS data are discrete with a very narrow range. Those who are not (e.g., *»year of birth«* does not have a narrow range) can safely be assumed to be, without any restriction.

b) the sample size is equal to all distributions, that is $n := n_{i_O} = n_{i_D} = n_{j_O} = n_{j_D} \; \forall i, j$. This holds, because subsampling the data is not an anonymization technique applied.

c) the domain X of possible values is the same for p_{i_O} and p_{i_D}, that is $X_{i_O} = X_{i_D}$. This might not be exactly accurate, as the idea of global recoding, for instance, actually *is* to alter the domain and combine codes. But, it turns out to be reliable enough to compute our information loss. Partially, the domains overlap, as some codes are identically transferred. If they do not, we take care for them to be mutually exclusive (i.e., one value may only exist in one domain).

d) missings are regarded as valid values. Although this seems strange at first, this accounts for our process of purging (see section 1.1), where missing values are partially preserved. As our anonymization process does not alter some missing values, there occurs no information reduction. The information *missing/not missing* is fully transferred from $\mathcal{I}_O$ to $\mathcal{I}_D$.

With this assumptions, one possibility to compare the distributions is calculating the Bhattacharyya coefficient (Bhattacharyya, 1946) for each variable v_i. This is defined as

$$\hat{w}_i := \sum_{x \in X} \sqrt{p_{i_O}(x) p_{i_D}(x)}$$

$\hat{w}_i$ is bounded between 0 and 1. It is 1 if the two distributions are identical, that is $p_{i_O} = p_{i_D}$, and 0 if absolutely no overlap between the two distributions exists, that is $p_{i_D} = 0$ if $p_{i_O} \neq 0$

and vice versa. Again, with this done, we can infer

$$\hat{\mathcal{I}}_E := \frac{1}{V} \sum_{i=1}^{V} \hat{w}_i$$

3. Results

Now we have a possibility to compute our weights, we apply this to all SUFs currently available – see table 1 for the corresponding DOI.

3.1 Onsite vs. Download

The results of comparing Onsite to Download are displayed in table 2. For all studies, both the total amount of variables as well as the amount of variables which are affected by anonymization are given. Those numbers are followed by the three estimators $\hat{\mathcal{I}}_P$, $\hat{\mathcal{I}}_H$, and $\hat{\mathcal{I}}_E$ we computed.

Table 2

Resulting values for the estimation of information loss in the Download version

	No. of variables				
	total	affected	$\hat{\mathcal{I}}_P$	$\hat{\mathcal{I}}_H$	$\hat{\mathcal{I}}_E$
SC1	2,168	661	.695	.696	.711
SC2	2,727	1,242	.545	.552	.594
SC3	3,743	1,025	.726	.733	.761
SC4	4,350	895	.794	.797	.829
SC5	2,573	425	.835	.839	.870
SC6	2,545	379	.851	.854	.895
BW	1,520	760	.500	.500	.518
TH	1,807	20	.989	.989	.996
Total	21,433	5,407	.748	.751	.780

As you can see, all three estimators of one Starting Cohort are similar to each other. This happens because the usual treatment for one variable results either in $\mathcal{I}_O(v_i) = \mathcal{I}_D(v_i)$ or $\mathcal{I}_D(v_i) = \emptyset$ (96% of all variables share this fate). Both of those outcomes are equally measured in all three estimators as $w_i = 1$, or $w_i = 0$, respectively.

Two studies stand out by very low values: SC2 with values between .54 and .59 and BW with an estimated information transfer between .50 and .52. This has two reasons:

First, those studies have a disproportionate amount of variables in *Institutions* (i.e., the data file containing information about the institution/school). This data file is completely removed in $\mathcal{I}_D$, so all variables produce full information loss. This are 547 variables (20% of all variables) in SC2 and 723 variables (48%) in BW. In SC3, SC4 and SC5, this data file is also present (and likewise removed), but here it only makes out 4% to 9%. The studies TH and SC6 do not have such datafile completely blocked in the Download version, which easily explains the high values of the corresponding estimators.

Second, most studies have additional data files available Onsite, containing micro data (i.e., regional context-information about the place of residence or institution), called *Microm* and *RegioInfas*. The structure of those datafiles is equal throughout the studies and contain the same 188 (Microm) or 68 (RegioInfas) variables. In school cohorts (i.e., SC2, SC3, SC4) this datafiles are available both for the place of residence as well as the place of the institution, doubling the variables whose information loss is counted. In the case of SC2, those 512 variables make out 19%.

As we regard this micro data as *additional* information available Onsite, rather than information removed in the Download version, we corrected the above estimators by removing those datafiles from the analysis. The results are given in the following table 3.

Table 3
Resulting values for the estimation of information loss in the Download version without regional data

	No. of variables				
	total	affected	$\hat{\mathcal{I}}_P$	$\hat{\mathcal{I}}_H$	$\hat{\mathcal{I}}_E$
SC1	1,980	473	.761	.762	.779
SC2	2,215	730	.670	.679	.731
SC3	3,231	513	.841	.849	.881
SC4	3,872	417	.892	.896	.931
SC5	2,402	254	.894	.899	.932
SC6	2,291	125	.945	.949	.995
BW	1,520	760	.500	.500	.518
TH	1,807	20	.989	.989	.996
Total	19,318	3,292	.830	.834	.865

3.2 Onsite vs. Remote

For the sake of completeness, we made the same calculations comparing the Onsite version to the Remote version. The results are given in table 4. Here we only focus on a comparison without regional data. As you can see, except for SC5, the first two estimators are equal, because the only method applied is purging string variables (which in both estimators is equally handled). In SC5, the datafile *Institutions* (containing 119 variables) is basically only available Onsite, but has eight variables made available in Remote, yet aggregated. This is the only time a different anonymization method has been applied.

Table 4
Resulting values for the estimation of information loss in the Remote version without regional data

	No. of variables				
	total	affected	$\hat{\mathcal{I}}_P^R$	$\hat{\mathcal{I}}_H^R$	$\hat{\mathcal{I}}_E^R$
SC1	1,980	50	.975	.975	.980
SC2	2,215	45	.980	.980	.985
SC3	3,231	32	.990	.990	.994
SC4	3,872	47	.988	.988	.994
SC5	2,402	184	.923	.924	.953
SC6	2,291	48	.979	.979	.999
BW	1,520	20	.987	.987	.987
TH	1,807	6	.997	.997	.997
Total	19,318	432	.978	.978	.987

4. Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS). From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

Bhattacharyya, A. (1946, July). On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics*, *7*(4), 401–406.

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

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P.-P. (2012). *Statistical Disclosure Control.* Chichester: Wiley.

Koberg, T. (2016). Disclosing the National Educational Panel Study. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study (in press).* Wiesbaden: Springer VS.

Oganian, A. (2011). Multiplicative Noise for Masking Numerical Microdata Data with Constraints. *SORT - Statistics and Operations Research Transactions (Special Issue)*, 99–112.

Skopek, J., Pink, S., & Bela, D. (2013). *Starting Cohort 4: 9th Grade (SC4). SUF Version 1.1.0. Data Manual.* National Educational Panel Study (NEPS), University of Bamberg.