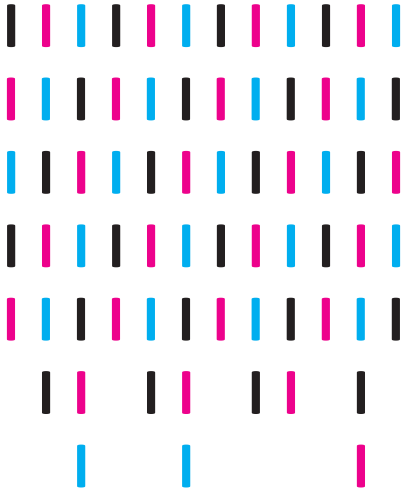

NEPS SURVEY PAPERS

Daniele Florean, Johannes Hofmann, and Corinna Kleinert

LONGITUDINAL CLASS IDENTIFIERS FOR NEPS STARTING COHORT 3: GENERATION PROCESS AND APPLICATION EXAMPLES

NEPS Survey Paper No. 52
Bamberg, March 2019

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

Longitudinal class identifiers for NEPS Starting Cohort 3: Generation process and application examples

Daniele Florean, Leibniz Institute for Educational Trajectories, University of Bamberg, University of Trento

Johannes Hofmann, Leibniz Institute for Educational Trajectories

Corinna Kleinert, Leibniz Institute for Educational Trajectories, University of Bamberg

E-mail address of lead author:

daniele.florean@stud.uni-bamberg.de

Bibliographic data:

Florean D., Hofmann J., & Kleinert, C. (2019). *Longitudinal class identifiers for NEPS Starting Cohort 3: Generation process and application examples* (NEPS Survey Paper No. 52). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP52:1.0

Longitudinal class identifiers for NEPS Starting Cohort 3: Generation process and application examples[1]

Abstract

In the German National Educational Panel Study (NEPS) Starting Cohort 3 (SC3) students in Grade 5 were selected within their school and class contexts. Up to now, class structures could be identified only cross-sectionally in the data and it was not possible to identify class membership longitudinally over the years. This survey paper describes the process of generating a new set of longitudinal class identifiers for the SC3 participants in school contexts from grade 5 to grade 9. The generated identifiers are consistent not only within but also across waves, allowing for longitudinal analyses of class membership and network effects. We then present a brief description of class mobility from grade 5 to grade 9 for the participants of the NEPS-SC3 samples drawn in 2010 and 2012.

Keywords

class group, class identifier, longitudinal dataset, peer effects, Starting Cohort 3

[1] We thank Anika Schenk-Fontaine for language revision.

1. School and class context data in NEPS Starting Cohort 3

Starting Cohort 3 (SC3)[2] (grade 5) is one of the six age groups (Starting Cohorts) sampled and followed by the German National Educational Panel Study (NEPS). The first interview of this Starting Cohort took place in autumn and winter 2010 (IEA DPC, 2011b). It is designed as a longitudinal study of students starting in grade 5 and following them within their class context in annual intervals in order to assess their general cognitive and curricular competences and to observe their educational careers, educational outcomes and development (IEA DPC, 2011b). The students were selected based on their schools, similar to the design of the SC2 und SC4 (IEA DPC, 2011a; 2011c), whereas the SC1 and SC6 rely on population-based individual sampling (IEA DPC, 2013; 2011d). In the German school system, grade 5 is normally considered to be the first grade in lower secondary education. There are a few exceptions to this rule, namely six-year primary schools in Berlin and Brandenburg. Students enrolled in these two federal states were also targeted in grade 5 to maintain comparability (IEA DPC, 2011b).

The initial sample of the NEPS-SC3 is composed of 365 schools throughout the 16 German federal states. Those schools included 214 secondary schools, 26 primary schools in Berlin and Brandenburg, 60 secondary schools with a high share of students with migration background to reach a higher representation for this group, and 65 schools for children with special needs *(Förderschulen).* In the first wave, the survey consisted of a maximum of two classes per school and included 5,525 students in total (IEA DPC, 2011b).

The student survey program consisted of cognitive assessments across different subjects and a paper-and-pencil questionnaire, which was filled out in class. Additional context information on students, classes, and schools was collected through parents, primary class teachers, German and math teachers, and school principals. Parents were interviewed via computer-assisted telephone interviews and teachers provided information about their class and their respective subjects via a postal survey (IEA DPC, 2011b).

In the NEPS-SC3 scientific use files, all students are assigned an individual identifier (ID_t), which is consistent over time, a school identifier (ID_i), which is consistent as long as the students remain at the school they joined in the first survey wave, and a class identifier (ID_cc), which is composed of the school identifier and a consecutive number starting with “01” to identify the class to which the students belong. This consecutive number is generated anew in each wave. Thus, it identifies students who are in the same class in a single wave but does not offer any information on the class composition within a given school over time (IEA DPC, 2016).

There are several reasons for why students may leave their original class context. In the following, we describe the three most important scenarios and their consequences for study design and data.

[2] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi:10.5157/NEPS:SC3:7.0.1. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network (Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J., 2011). The aim of the study is to collect longitudinal data on educational decisions, competences and lifelong returns on education for the German population.

- **Scenario 1:** A student reaches the next grade but is transferred to another class in the same school due to administrative reasons at the school level (e.g., to balance out the number of students per class or to re-distribute students after choosing a second foreign language). In this case, the student will receive a new class identifier (ID_cc), but stays part of the main survey field within the school. In order to document this type of within-grade and within-school move, the NEPS-SC3 class identifier is generated anew at each panel wave based on the within-school class structure in the given data collection period. Due to the fact that NEPS-SC3 respondents are increasingly intermingled in classes that have not been part of the survey previously, the number of class identifiers increases as the survey proceeds. The new classes are monitored by the teacher questionnaires and receive their own class identifiers each year, as long as a least one NEPS-SC3 participant is part of them. Students in these classes who have not been part of the survey in the first wave will not be integrated retrospectively and are therefore not included in NEPS data (IEA DPC, 2016).

- **Scenario 2:** A student has to repeat a grade due to low performance or skips a grade due to high performance. In this case the student remains in the same school and therefore keeps his/her school identifier. However, he/she **enters** a group called "individualized main field" and is not assigned a new class identifier, but a missing value in ID_cc. Those students can still be reached through the school and receive the same survey program as the main field but are irrelevant for keeping track of class change because they left the grade in focus (IEA DPC, 2016).

- **Scenario 3:** It is not possible to reach the student through the school anymore. This may be due to a number of different reasons: First, the student has left his/her school for individual reasons. Second, different cases can be identified where whole schools cease to exist in the data. Schools are no longer observed in the NEPS when the number of target persons per school has dropped below three students, when they refuse to participate in the survey any longer, when they do not offer the grade in focus (for example primary schools in Berlin and Brandenburg after grade 6), or when they have been closed down. All students in scenario 3 are followed up individually (mostly via postal surveys), which means that the school and the class identifiers are both dropped and only the individual identifier is kept. Respondents in this group are, like the students in scenario 2, irrelevant for keeping track of class changes (IEA DPC, 2016).

To compensate for the loss of students in the main field (especially after grade 6 in Berlin and Brandenburg due to the end of the 6-year primary schools), the NEPS-SC3 sample was augmented in wave 3 by 2,205 grade 7 students. Those students received the same three identifiers (ID_t, ID_i and ID_cc) as the other students did in grade 5. However, class changes in grade 5 and grade 6 cannot be identified retrospectively for them (IEA DPC, 2016).

All in all, the sample size of the main survey field (with all three identifiers present) declined from 5,525 students in wave 1 to 5,174 students in wave 5 (with a spike of 6,452 students in wave 3). The individualized main field containing students that repeated or skipped a grade within the same school included 121 students in wave 5, and the group which was tracked individually grew from 355 students in wave 2 to 2,108 students in wave 5 (IEA DPC, 2016).

In wave 6 (grade 10), the class identifier was not generated at all. After grade 9, lower secondary education ends and students from the lowest educational track *(Hauptschule)* may leave school. For the NEPS-SC3 sample this resulted in a significant number of students receiving individual follow-up questionnaires. Therefore, we will generate longitudinal information which make the tracking of class changes throughout the first five survey waves, from grade 5 to grade 9, possible.

2. From cross-sectional to longitudinal class identification

Since the late 1960s, it is well known in educational research that characteristics and achievements of the students' peers have an effect on the students' own achievements (Coleman 1966). We can distinguish two main streams of research on peer effects in education. On the one hand, there are studies focusing on the reciprocal influence of students' and classmates' educational achievement (e.g. Slavin 1990; Sacerdote 2011; Hanushek et al. 2003; Lavy et al. 2012). On the other hand, studies examine how average characteristics of the peer group, such as social origin, race, or immigration status, influence the educational achievement of students (among others, van der Slik, Frans W. P. et al. 2006; Stewart 2007; Opdenakker and van Damme 2007; Caldas and Bankston 1997; Agirdag et al. 2012). Some researchers recommend a longitudinal approach in order to avoid multicollinearity and reflexivity problems (Hanushek et al. 2003), while the use of fixed effects at student, class, and school levels is recommended to avoid problems of self-selection (Sacerdote 2011).

In order to implement these methodological refinements, a reliable and coherent way to identify class membership throughout secondary schooling is needed. Moreover, a longitudinal class identifier is useful to investigate how movement between classes (e.g., changing classes between grades or the introduction of a new classmate in a group) influences the achievement of both individuals and groups. Additionally, implementing a class identifier that is consistent between waves enables researchers to conduct longitudinal network-based research. For example, being able to identify students who share the same class over time and, therefore, identify class composition changes allows researchers to investigate how changing class membership and meeting new peers affects educational achievement or how the arrival of new students affects class performance.

The class identifier (ID_cc) available in the NEPS Starting Cohort 3 data, however, is only useful for cross-sectional research. It always resembles the same pattern, including the school identifier and a consecutive number. Receiving the same class identifier for two or more years in a row does not mean that the student stayed in the same class. It only signifies that the student is situated in the same class context as other students with the same class identifier within a given school year. Therefore, the available class identifier cannot be used for tracking class changes. In order to track the pathway individuals take through lower secondary education (within their school of origin), it is, therefore, necessary to implement a class identifier which consistently identifies class memberships over different survey waves.

3. Rationale and aims of new identifiers

In order to reach this aim, we developed an approach that addresses the lack of consistency between class identifiers over time and built a Stata syntax which generates new, longitudinal class identifiers for the first five survey waves (from grade 5 to grade 9) of the NEPS-SC3 in long and wide format.

To generate the new class codes, we proceeded on two basic assumptions: First, even if the class identifiers available in a given wave may have changed, the largest subgroup in each wave should "carry on" the identifier the respondents belonging to it shared in the wave before. The underlying idea is that, for each class group, the students who change class or drop out each year are fewer than the students who remain together in the same class. Second, we assumed that the existing class codes within a given school and wave are consistent and valid. In other words, the fact that students who share the same class code in a given wave means that they were indeed in the same class at that time, even if they did belong to different classes in the first wave of observation.

As an example to better understand the difference of the original and new class identifiers, two tables are provided. Table 1 shows a (purely fictional) example school as it would appear in the NEPS-SC3 *CohortProfile* dataset in wide data format, ordered by the respondent identifier. In fictional school 10012, 16 students were surveyed in grade 5, 11 in class 1 and another 5 in class 2. It can be easily seen that the original class identifier codes change between waves following no specific rationale. Therefore, at least at first glimpse, no clear class structure over time is visible and it is not possible to identify and follow students belonging to the same class over time based on these codes.

Table 2 shows the same school with the new class identifiers resulting from the recoding process, ordered again by the respondent identifier. The class structure is significantly clearer now and movements between classes are more easily identifiable.

As can be seen from Table 1 and 2, respondents who started grade 5 in class 1 (respondents 123-133) are split in two different classes in grade 6. To reliably identify consistent class groups, we assigned class code 1 to the larger of the two subgroups, which had originally been assigned class code 3 in grade 6. In a similar fashion, for the respondents who had been in class 2 in grade 5 (respondents 134-138), the original class code in grade 6 was replaced with the new class code 2. Given that, in grade 6, this class code was shared with respondents 123-125, who had originally been in class 1 in grade 5, they were also assigned the new class code 2 to reflect the fact that these persons changed from class 1 to class 2. In grade 7, both members of class groups 1 and 2 in grade 6 carry on their new codes, since there are no changes in class composition visible. Only respondent 126 moves to a different class, which therefore is assigned the new code 3. These three class codes are carried on consistently for wave 4 and wave 5, since there are no further changes in the class composition of the example school, except for occasional school dropouts.

Table 1

Example school with original class IDs

ID School	Id Person	ID Grade 5	ID Grade 6	ID Grade 7	ID Grade 8	ID Grade 9
10012	123	1	1	3	1	1
10012	124	1	1	3	1	-
10012	125	1	1	3	1	1
10012	126	1	3	1	2	2
10012	127	1	3	2	3	3
10012	128	1	3	2	3	3
10012	129	1	3	2	3	-
10012	130	1	3	2	3	3
10012	131	1	3	2	3	3
10012	132	1	3	2	-	-
10012	133	1	3	-	-	-
10012	134	2	-	-	-	-
10012	135	2	1	3	1	1
10012	136	2	1	3	1	1
10012	137	2	1	-	-	-
10012	138	2	1	3	-	-

Table 2

Example school with new longitudinal class IDs

ID School	Id Person	ID Grade 5	ID Grade 6	ID Grade 7	ID Grade 8	ID Grade 9
10012	123	1	2	2	2	2
10012	124	1	2	2	2	-
10012	125	1	2	2	2	2
10012	126	1	1	3	3	3
10012	127	1	1	1	1	1
10012	128	1	1	1	1	1
10012	129	1	1	1	1	-
10012	130	1	1	1	1	1
10012	131	1	1	1	1	1
10012	132	1	1	1	-	-
10012	133	1	1	-	-	-
10012	134	2	-	-	-	-
10012	135	2	2	2	2	2
10012	136	2	2	2	2	2
10012	137	2	2	-	-	-
10012	138	2	2	2	-	-

4. Generation process for the SC3

Below we describe in detail how we generated the new longitudinal class identifiers. The actual generation process was performed using Stata (Version 14). Stata syntax to replicate this process is provided for download together with this NEPS survey paper on the NEPS and

LIfBi website. For generating the new measures, we used data from the NEPS-SC3 scientific use file 7.0.1 download version (doi:10.5157/NEPS:SC3:7.0.1)[3].

The **first step** was to carefully examine the available data and restrict it to the relevant population. Survey participants in special needs education schools were excluded, due to the peculiar class structure of those schools. Participants in elementary schools from the Berlin-Brandenburg region were excluded as well, due to the fact that classes were introduced (but not observed in NEPS-SC6) in grade 1 and exist only until grade 6. Data from the sixth survey wave onwards were not taken into account due to the already mentioned lack of class identifiers in grade 10.

To make it easier to work with the old identifiers, some preliminary data preparation was necessary. First, we generated a variable which contains only the last two digits of ID_cc, which denote the consecutive number of the class. Since these two digits never exceeded "09", only the last digit of the original class code was kept and converted to string format. Then the dataset was reshaped into wide format, keeping only the school identifier (ID_i), the respondent identifier (ID_t), and the reduced class identifier (string_ccN, where N = [1,5] is the wave identifier).

In order to identify groups of students who are consistently together in one class up to any given wave, we combined the string identifiers string_ccN into a "trajectory" string variable for each person and wave, which uniquely identifies the class groups each respondent belonged to in all the survey waves up to the current one. In this variable, subgroups of respondents who remained in the same class together until a given wave are characterized by sharing the same trajectory. This means, it carries clear information on the class group each respondent joined in each wave, as well as on the class group each respondent changed to in the subsequent wave. Subgroups will be, from this point on, defined by sharing a trajectory and not a particular class code, since class codes do not systematically change between waves (and it's impossible ,for example, to be sure that the class code "3" for a given school at wave 4 identifies the same group of persons it identified in the previous wave). It is actually the use of trajectories that makes it possible to define new consistent class identifiers. Those "subgroups" will receive the new longitudinal class identifiers in the end.

As an example, Table 3 shows again the fictional school from the previous paragraph and the trajectories of all the students in this school, which were derived by combining the consecutive numbers of the original class identifiers into a single string variable. The value "0" replaces missing values and thus clearly denotes waves in which students left the main survey field (i.e. their original grade or school). In Table 3, we can see, for example, that all students identified by class code "2" in grade 7 (ID person 127 to 132) are identified by class code "3" in the previous grade.

[3] Generation and tests of the new class identifiers with syntax provided were conducted on the indicated edition of the data. Different data editions could require changes in the provided syntax.

Table 3

Fictional example school with class trajectories

ID School	Id Person	traj. Gr. 5	Traj. Gr. 6	Traj. Gr. 7	Traj. Gr. 8	Traj. Gr. 9
10012	123	1	11	113	1131	11311
10012	124	1	11	113	1131	11310
10012	125	1	11	113	1131	11311
10012	126	1	13	131	1312	13122
10012	127	1	13	132	1323	13233
10012	128	1	13	132	1323	13233
10012	129	1	13	132	1323	13230
10012	130	1	13	132	1323	13233
10012	131	1	13	132	1323	13233
10012	132	1	13	132	1320	13200
10012	133	1	13	130	1300	13000
10012	134	2	20	210	2100	21000
10012	135	2	21	213	2131	21311
10012	136	2	21	213	2131	21311
10012	137	2	21	210	2100	21000
10012	138	2	21	213	2130	21300

The **second step** of the overall generation process was to generate the additional information necessary to generate the new longitudinal class identifiers. We first built a variable that identified the largest subgroup of students in a given school that shares a common trajectory starting in wave 1 (grade 5) until the wave taken into account (max_N_c, where N is the wave number from 2 to 5). The rationale behind this variable is that, as explained in the previous section, the largest subgroup for each starting class should carry on in all following waves the original class code (*1* or *2*) that was assigned in the first wave of observation.

As an example, in Table 3 for starting class 1 in grade 5, the largest subgroup in grade 6 is characterized by the trajectory “13”. Therefore, this subgroup will carry on the original code (“1”). Consequently, class code “3” in grade 6 will be replaced by the new class code “1”. Moving on to grade 7, the largest subgroup that shares the trajectory “13” in the previous wave is subgroup “132”. Therefore, this group will carry on the new class code “1” that will replace the original class code “2” in grade 7 and so on.

Some schools posed the additional challenge of having subgroups of the same size in some waves. In order to deal with this problem, we generated a variable that counts (for each starting class group, in each school) the number of largest subgroups of the same size in each wave (samec_N). The variable takes missing values if the subgroup considered is not the largest in the wave for that starting class and school.

In the **third step**, the new class identifiers were generated. This process follows the same steps both for respondents sampled in 2010 (Wave 1) and for respondents sampled in 2012 (Wave 3). Longitudinal class identifiers for the respondents belonging to the two different samples were generated separately, and the resulting two datasets are appended at the end of the generation process. Below we describe the process for the subsample drawn in 2010 (Wave 1).

First, we generated a new variable for the first wave class identifier (new1) that takes ordered values, differently from the original starting class identifier, which sometimes "jumps". By design, each school in the sample should consist of only two sampled classes. Therefore, in theory, the variable *new1* should take only values 1 and 2. Due to oversampling of students with migration background, for a small number of cases (N=71) the value of this variable goes up to 5. These cases were treated in the same way as "regular" cases in the following process. It will be up to the researcher to decide whether to include those cases in their analysis or not.

Then, for each wave after the first (N = [2,5]), the algorithm to generate the new class codes proceeds in the following steps:

1. Generate the new class identifier (newN) with only missing values (.)
2. Sort the respondents by school identifier and starting class, and within school identifier and starting class, by frequency of the subgroup code (variable t_N_c) and subgroup code within frequency (variable t_N) for each wave up to the Nth
3. Replace the new class identifier (newN), missing, with the previous wave's new class identifier (newN-1), effectively carrying on the first wave identifier if the subgroup is the largest one and not problematic.
4. *Missing* class identifiers are at this point assigned (variable newN = 0), drawing this information from the original class identifier (stringN = 0).
5. Data are then sorted by the given wave's original class identifier (variable stringN) and, within that, by the new class identifier (variable newN). Within each "old" class identifier, missing values for each new class identifier are filled to ensure consistency between starting classes.
6. Still missing (.) new class identifiers are then filled progressively with a new code, and the previous step is repeated to ensure consistency between the two starting classes for each school. A control variable z_N is used to identify newly filled values in each wave in order to exclude them from step 2 and 3 for the following wave.
7. The process is repeated for the following wave, up to wave 5.
8. At the end of the whole cyclic process, the missing information indicator for the new class identifiers was replaced with the usual "." instead of "0."

For the subsample of respondents sampled in wave 3, the process is the same, with the only difference being that it starts with wave 3 and proceeds only for the following two waves.

Information on the new class codes is stored in two datasets including school identifier and respondent identifier and the new class code for each wave, which is available both in long and wide format to ensure the maximum usability and compatibility with the rest of the NEPS SC3 datasets. The variables "newN" are renamed "ID_class_N" for the wide format dataset and "ID_class" for the long format dataset. A "wave" variable ensures compatibility of the dataset in long format with other long format datasets in the NEPS-SC3 SUF.

5. Applying the algorithm to the SC2 and SC4

The process described above should, in theory, be applicable to all NEPS school samples that share the same ID structure as the SC3 for school and class to which the respondent belongs, namely Starting Cohort 2 and Starting Cohort 4. In the following, we describe problems and processes that turned up in practice.

5.1 Starting Cohort 2

Respondents in the SC2 were selected when they were in Kindergarten (wave 1) or in grade 1 in elementary school (wave 3). Class identifiers for Starting Cohort 2 are hence used in waves 3 to 6 to follow respondents' class affiliation through elementary school (grades 1 to 4). As in the SC3, in the SC2 waves 3 to 6, a school identifier (ID_i) and a class identifier (ID_cc) are available in the SUF.

The only difference in the data structure between the SC2 and the SC3 is that the class identifier for grade 1 (wave 3) exceeds 9 in rare cases, with some schools having up to 12 different class groups in grade 1. In order to keep the same trajectory structure as in the SC3, we replaced the two-digit class identifiers with a letter, which, in the string structure used by Stata, occupies the same space as a single digit. Since the subsequent process also assigns a numerical identifier to those cases when building the new class codes dataset, the replacement is temporary and does not impact the final result.

Since SC2 respondents are sampled at the beginning of kindergarten and at the beginning of elementary school, we only need one syntax for generating new class identifiers for elementary school years, differently from SC3 where we needed two syntaxes, one for each sub-sample.

The process of generation of the new class IDs faithfully follows the process used for the SC3. It begins generating the new wide-format "trajectory" dataset which is at the base of the whole process, with the above-mentioned exception for respondents whose class codes go above 9 in order to keep the same number of characters in the string variable. Therefore, "10" becomes "a", "11" becomes "b" and "12" becomes "c". There were no cases where the class code took values greater than "12".

After the "trajectory" dataset is built, and after the same variables that are necessary for the process in the SC3 are generated (largest subgroup, largest same-size subgroup), the new class identifiers are generated following the same process used for the SC3.

First, the class identifier for wave 3 (grade 1) is re-generated so that it takes "continuous" values, that is, without "jumps," within each school. Then, as in the SC3, for each school, in each subsequent wave (waves 4 to 6), the largest subgroup (i.e., group of people that shares a common trajectory up to the considered wave) within each new class group "carries on" the initial class code, while the other groups of students sharing different common trajectories get a progressive class identifier (for a step-by-step description of the process refer to section 4 of this paper). In the end, the new class identifier variables are renamed and re-labeled according to the wave and grade they refer to, and both long and wide format dataset

containing the new class identifiers are generated. A "wave" variable ensures compatibility of the long format dataset with the other datasets of the SC2 SUF family[4].

The Stata syntax we used for the generation of longitudinal class identifiers for the SC2 is available for download as well.

5.2 Starting Cohort 4

In contrast to the SC2, NEPS-SC4 presents a whole series of issues that, in the end, made us desist from generating new class code identifiers. Researchers who intend to use longitudinal class identifiers for their analyses despite these problems may adapt our syntax on their own to this starting cohort.

Starting Cohort 4 data follows students in secondary schools from grade 9 onwards. At least traditionally, this is the last grade of *Hauptschule* in the German educational system, whereas in reality today many Hauptschule students stay in school for another year. Similarly, *Realschule* also lasts for an additional year (until grade 10). *Gymnasium* students are the only students who stay in secondary school until grade 12 or grade 13. Moreover, the last two grades in *Gymnasium* usually do not follow the classical class structure anymore. Instead, students may choose a set of different subject-centered courses consisting of different co-students, which could include any students in the same grade. This further lowers the usefulness of a longitudinal class indicator.

Accordingly, data investigation shows that valid class identifiers are available in NEPS-SC4 only for wave 1 (Fall 2010), 3 (2011/2012) and 5 (2012/2013) (i.e., grades 9-11). On top of that, for 37 of 149 *Gymnasien* in the sample, no valid class identifier is available for Wave 3, further reducing the usefulness of this indicator. Due to these peculiarities and data issues, we decided not to provide syntax for SC4.

6. Class Mobility and Stability

We tested the usefulness of the new class identifiers by looking at the mobility of respondents among classes, in order to see how many respondents experience a change in class identifiers over the five-year period of secondary schooling (or respectively, the three-year period for respondents sampled in wave 3).

To this end, we merged the new dataset *SC3_New_class_codes_long* with the *SC3_CohortProfile* dataset, version 7.0.1. We then dropped the non-merged respondents (students in Berlin/Brandenburg primary schools, students in special education schools) and students from the migration background oversampling, which contains about 70 cases. The process resulted in a total subsample of 7,204 respondents, 5,009 of which were sampled in wave 1 and 2,205 were sampled in wave 3, on which we conducted our brief analysis.

The following graphs show the rates of year-by-year mobility (Fig. 1 and 3) and overall mobility (Fig. 2 and 4) of our respondents. By year-by-year mobility we mean a change in class code or

[4] The process of generation of new class identifiers was conducted and tested on the NEPS SC2 SUF version D_7_0_0 (doi *10.5157/NEPS:SC2:7.0.0*). Different data editions could require some changes in the syntax provided.

school code compared to the previous year, while by overall mobility we mean having ever experienced a change in class code since the first wave of observation.

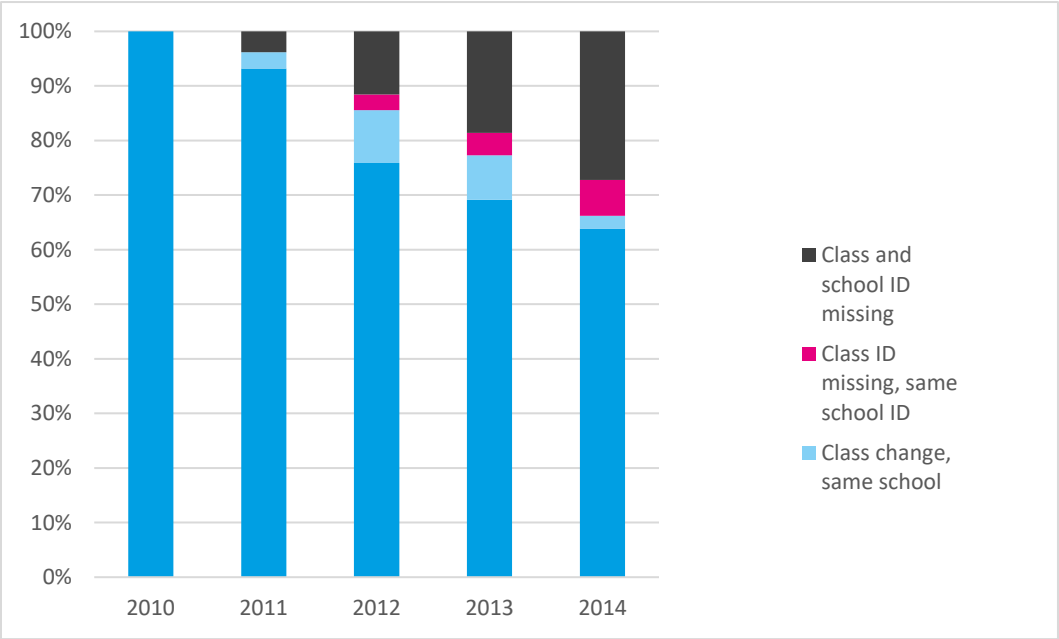

Figure 1. Year-by-year school and class mobility for students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1).

As described in the initial section (page 4), respondents who either leave their starting school or have to repeat a school year, exit from the main sample (but not the survey) and are tracked individually. Students that drop out of the main sample while remaining individually tracked (as described by scenario 3 in section 1, p. 4) are represented in black. The small percentage of students who loses their class identifier while maintaining their school identifier (between 2 and 6 per cent) is represented in purple and is generated by the latter case (as described in section 1, p. 4, scenario 2).

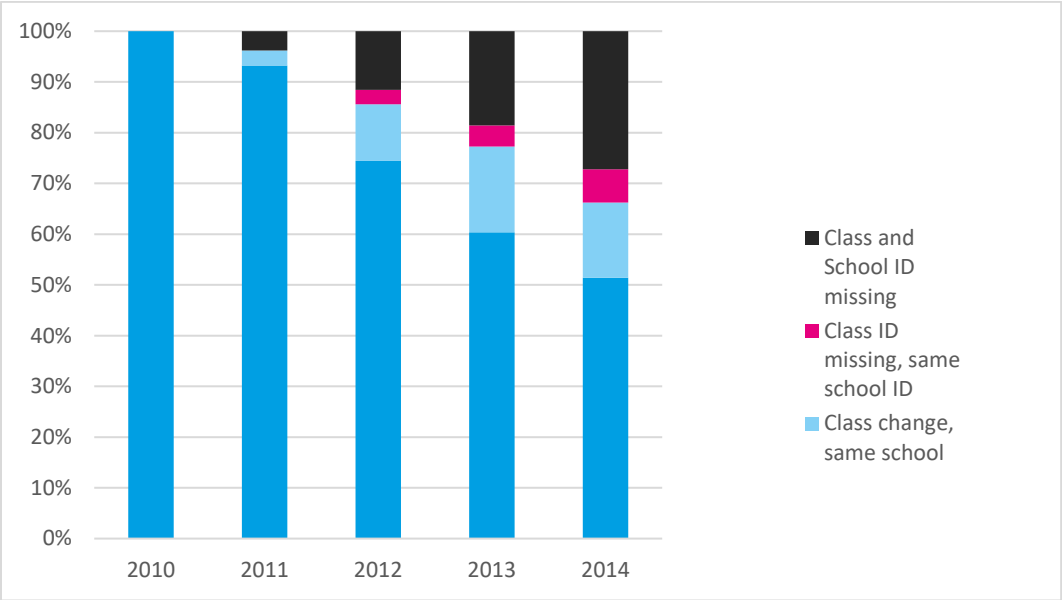

Figure 2. General school and class mobility of students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

Figure 1 shows that year-by-year mobility is higher in grade 7 and 8 than in grade 6, with 10% and 8% of the sample respectively changing class codes compared to the previous year. Changes between classes seem to be concentrated between waves 2 and 3 and between waves 3 and 4. Contrasted with this, the loss of class ID due to class repetition is concentrated between waves 4 and 5. What is more notable than within-grade, within-school mobility is the increasing number of students exiting their starting school and class either because they drop out or because they have to repeat the year. This is more visible in Figure 2, which shows general mobility: More than 25% of students is followed individually at wave 5. At the same time, if we focus on permanence within the class group, only slightly more than 50% of students in the sample reach wave 5 in the same class group they began their secondary school career. By wave 3, 12% of the respondents had changed class at least once. This amount increases to over 16% in the wave 4, to then decrease to almost 15% in wave 5. This counter-intuitive decrease is due to the fact that some of the respondents who changed class, dropped out of the sample or are followed individually. Therefore information on their class of belonging is lost.

It should be noted, however, that, because the largest subgroup in each wave carries on the code from the previous wave, the new coding system is inherently conservative, especially when looking at mobility rates in comparison to other countries or contexts.

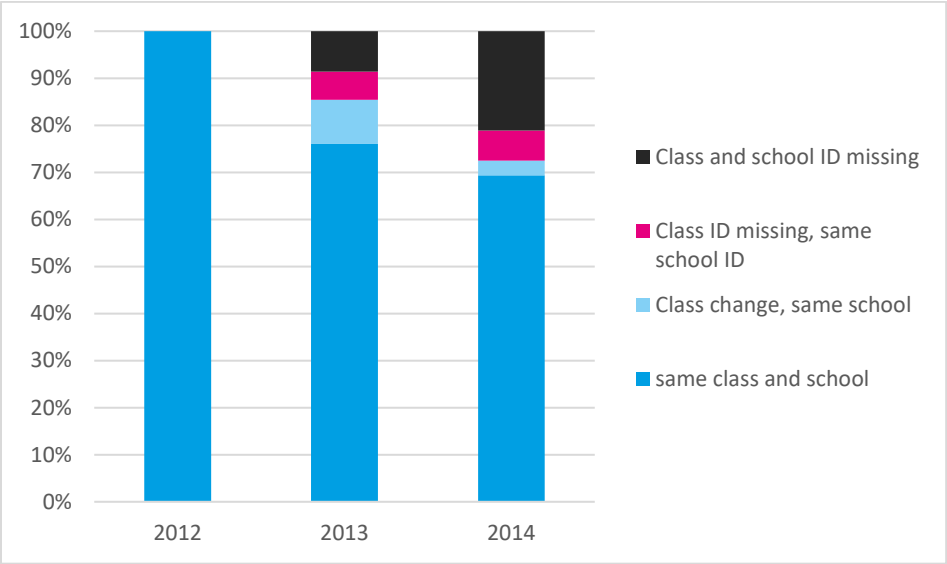

Figure 3. Year-by-year school and class mobility of students sampled in 2012. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

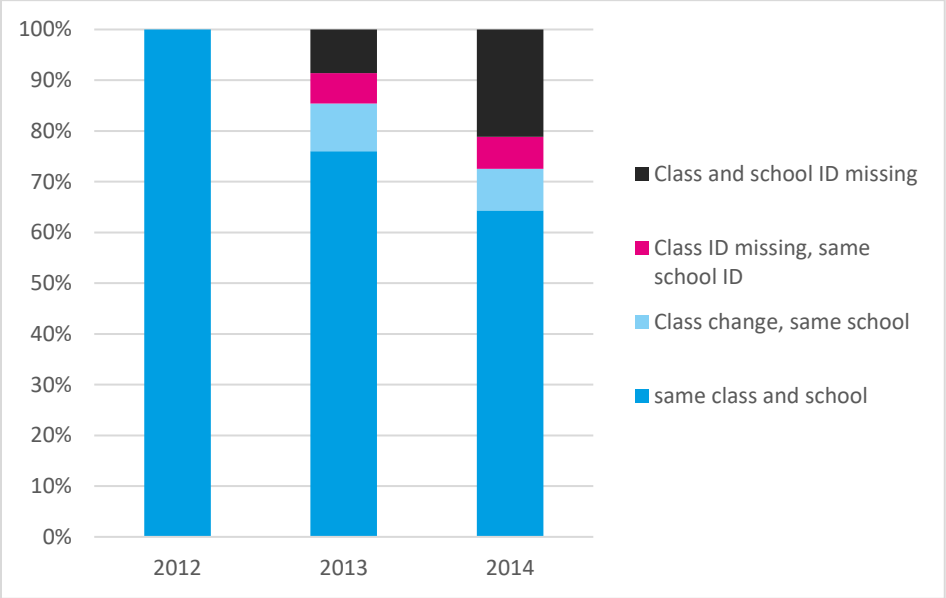

Figure 4. General school and class mobility of students sampled in 2012. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

Results of the analysis for the respondents sampled in wave 3 are shown in Figure 3 and 4. They mirror the results for respondents sampled in wave 1, with minor differences due to the shorter period of observation. Not surprisingly, year-to-year mobility is higher in wave 4 than in wave 5, but it is striking that there is a more than ten percentage point increase in respondents with class and school identifiers missing. Looking at general mobility (Figure 4), less than 65% of the respondents stays in the same class for all three years of observation.

These graphs present only an average picture of mobility between class groups and do not account for differences in mobility between school types, which are quite large. More detailed

information for each school type and subsample is available in Appendix 2 for the NEPS-SC3 students sampled in 2010.

7. Summary and conclusion

The class identifier available in the Starting Cohort 3 scientific use files does not account for class consistency across waves. Therefore, we developed a new set of longitudinal class identifiers that is consistent both within and across waves, allowing for their use in longitudinal data analysis. We then successfully extended the algorithm for generating longitudinal class identifiers to NEPS Starting Cohort 2 data on class structures in elementary schools. With the newly generated codes, we then examined between-class and between-school mobility of the students observed in NEPS Starting Cohort 3. By the end of the five-year observation period ranging from grade 5 to grade 9, only 50% of the 2010 sample and a little less than 65% of the 2012 sample do not experience any kind of mobility. However, mobility between classes within the same grades and schools is quite low, particularly when compared to between-school mobility.

Our brief analysis already shows that the new longitudinal class identifiers we provide with this NEPS survey paper may open new and interesting opportunities for answering longitudinal research questions using the data of the NEPS school samples, in particular for the study of peer and network effects.

References

Agirdag, O.; van Houtte, M.; van Avermaet, P. (2012): Why Does the Ethnic and Socio-economic Composition of Schools Influence Math Achievement? The Role of Sense of Futility and Futility Culture. In *European Sociological Review* 28 (3), pp. 366–378. DOI: 10.1093/esr/jcq070.

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

Caldas, Stephen J.; Bankston, Carl, III (1997): Effect of School Population Socioeconomic Status on Individual Academic Achievement. In *Journa of Educational Research* 90 (5).

Coleman, James S. (1966): Equality of Educational Opportunity. U.S. Department of Health, Education and Welfare.

Hanushek, Eric A.; Kain, John F.; Markman, Jacob M.; Rivkin, Steven G. (2003): Does peer ability affect student achievement? In *J. Appl. Econ.* 18 (5), pp. 527–544. DOI: 10.1002/jae.741.

IEA Data Processing and Research Center (IEA DPC) (2011a). Methodenbericht (Field Report) NEPS-Startkohorte 2. Haupterhebung – Winter/Frühjahr/Sommer 2011. A12. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/NEPS_FieldReport_SC2_W1_PAPI.pdf

IEA Data Processing and Research Center (IEA DPC) (2011b). Methodenbericht (Field Report) NEPS-Startkohorte 3. Haupterhebung – Herbst/Winter 2010. A28. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/1-0-0/Methodenbericht_SC3_W1_PAPI.pdf

IEA Data Processing and Research Center (IEA DPC) (2011c). Methodenbericht (Field Report) NEPS-Startkohorte 4. Haupterhebung – Herbst/Winter 2010. A46, A67, A83. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/Methodenbericht_A46_A67_A83.pdf

IEA Data Processing and Research Center (IEA DPC) (2011d). Methodenbericht (Field Report) NEPS Etappe 8. Befragung von Erwachsenen. Haupterhebung 1. Welle 2009/2010. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/1-0-0/Methodenbericht_SC6_W2_B72.pdf

IEA Data Processing and Research Center (IEA DPC) (2013). Methodenbericht (Field Report) NEPS-Startkohorte 1. Haupterhebung 2012/2013. B04. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC1/1-0-0/SC1_MB_1.pdf

IEA Data Processing and Research Center (IEA DPC) (2016). Methodenbericht (Field Report) NEPS-Startkohorte 3. Haupterhebung – Frühjahr 2015. A98. Available online (only in German): https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/6-0-0/Methodenbericht_SC3_W6_PAPI.pdf

Lavy, Victor; Silva, Olmo; Weinhardth, Felix (2012): The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. In *Journal of Labor Economics* 30 (2).

Opdenakker, Marie-Christine; van Damme, Jan (2007): Do school context, student composition and school leadership affect school practice and outcomes in secondary education? In *British Educational Research Journal* 33 (2), pp. 179–206. DOI: 10.1080/01411920701208233.

Sacerdote, Bruce (2011): Peer Effects in Education. How Might They Work, How Big Are They and How Much Do We Know Thus Far? In, vol. 3: Elsevier (Handbook of the Economics of Education), pp. 249–277.

Slavin, Robert E. (1990): Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis. In *Review of Educational Research* 60 (3).

Stewart, Endya B. (2007): School Structural Characteristics, Student Effort, Peer Associations, and Parental Involvement. In *Education and Urban Society* 40 (2), pp. 179–204. DOI: 10.1177/0013124507304167.

van der Slik, Frans W. P.; Driesse, Geert W. J. M.; De BOt, Kees J. L. (2006): Ethnic and Socioeconomic Class Composition and Language Proficiency. A Longitudinal Multilevel Examination in Dutch Elementary Schools. In *European Sociological Review* 22 (3), pp. 293–308. DOI: 10.1093/esr/jci058.

Appendix 1: General and year-by-year mobility of students sampled in 2010, by school type

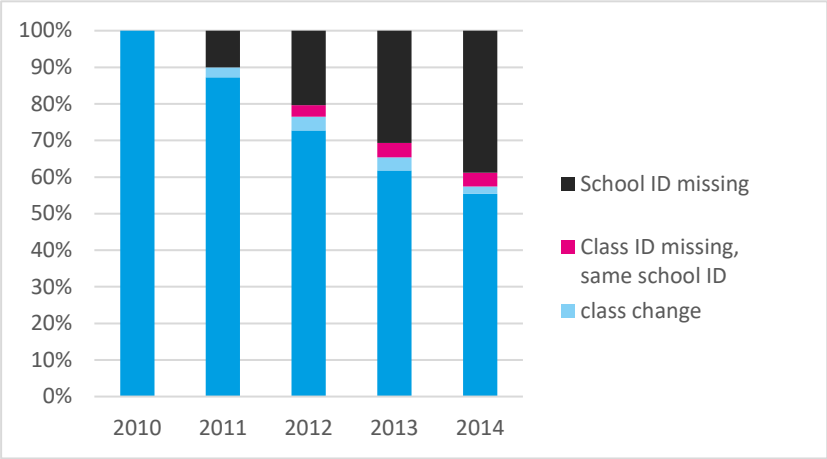

Figure 5. Year-by-year mobility for Hauptschule students sampled in 2010. Authors' own elaboration on NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

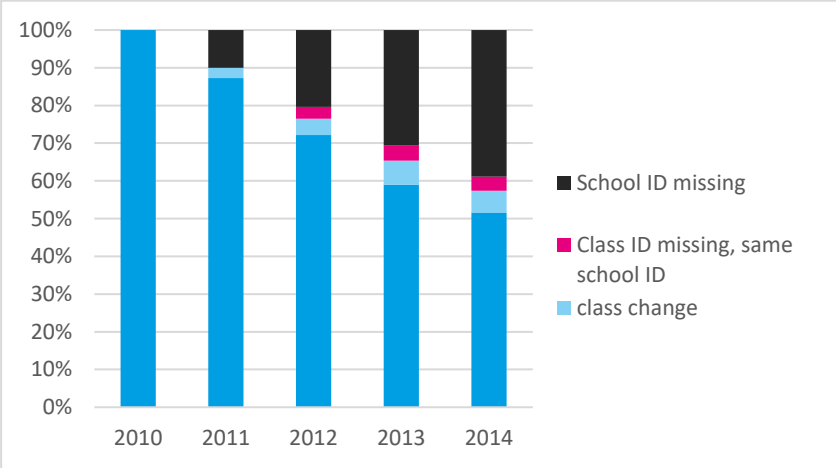

Figure 6. Overall mobility for Hauptschule students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

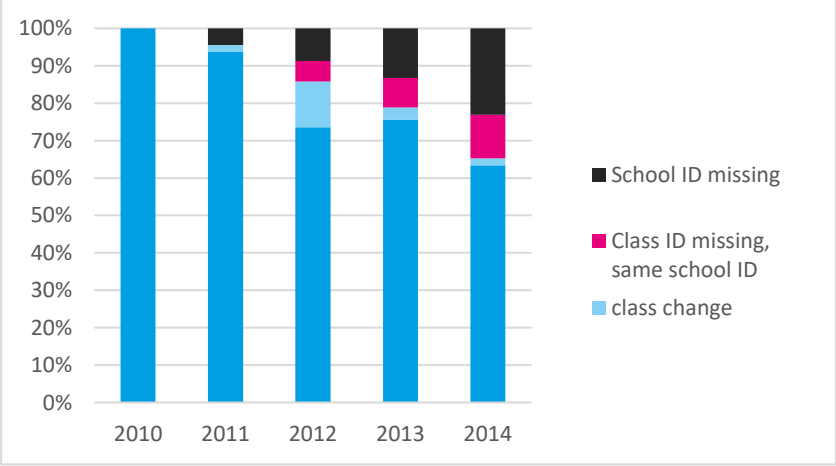

Figure 7. Year-by-year mobility for Realschule students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

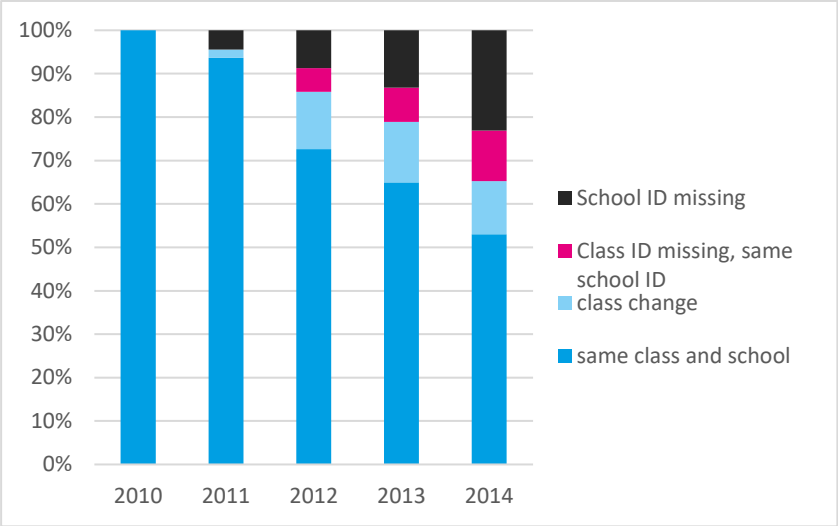

Figure 8. Overall mobility of Realschule students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

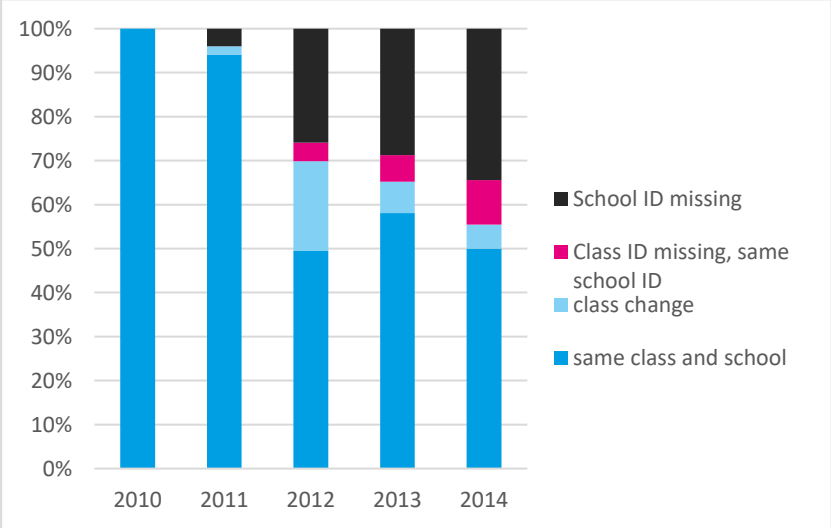

Figure 9. Year by year mobility of students from schools with multiple education courses sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

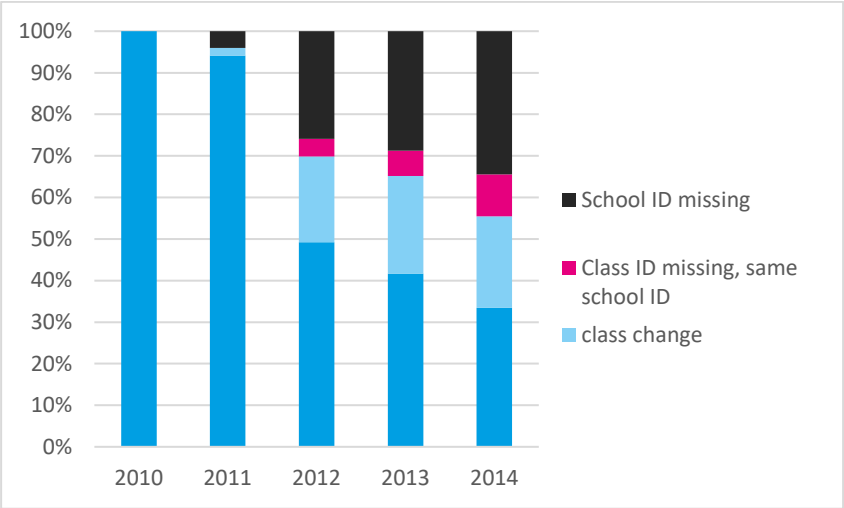

Figure 10. Year by year mobility of students from schools with multiple education courses sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

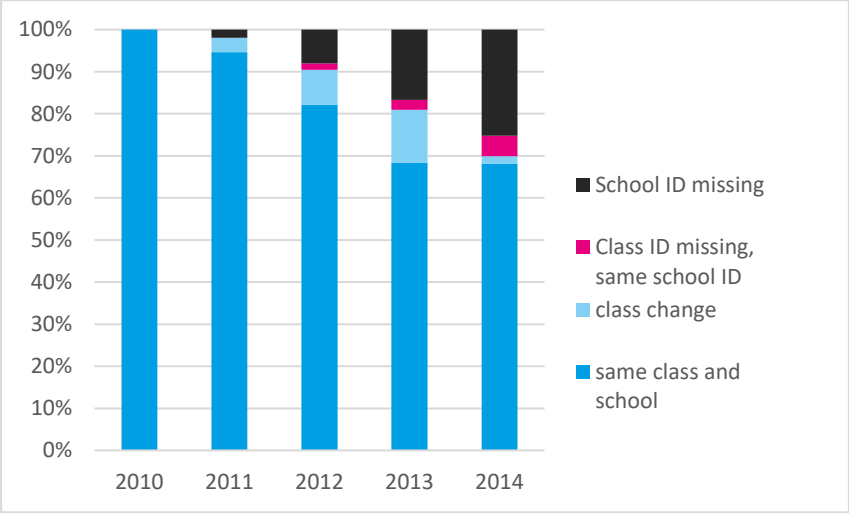

Figure 11: Year by year mobility of Gymnasium students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

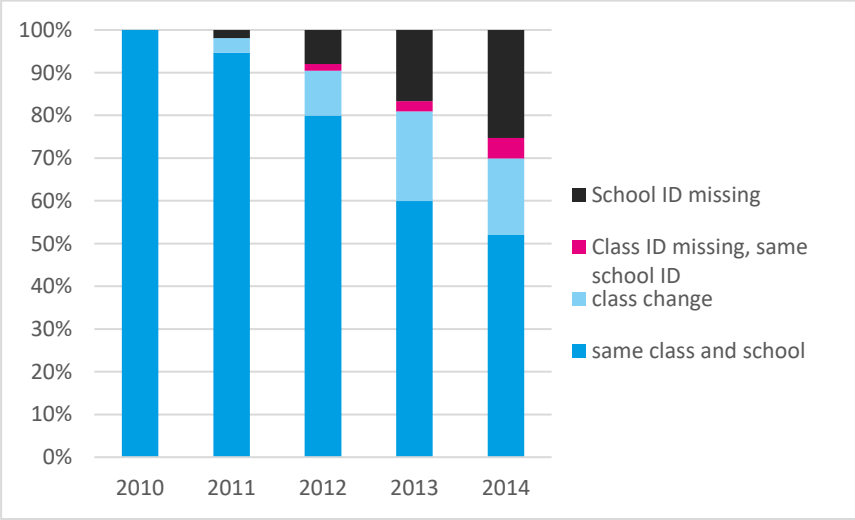

Figure 12: Overall mobility of Gymnasium students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

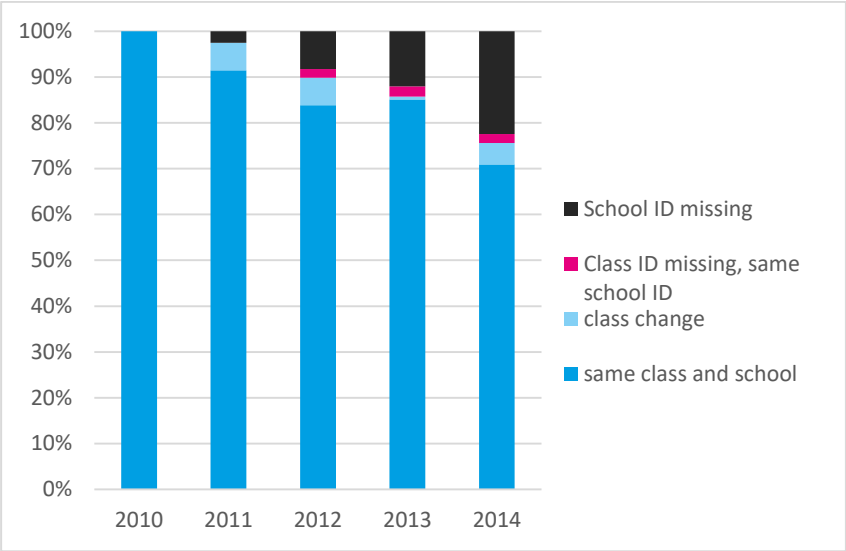

Figure 13. Year by year mobility of Comprehensive schools students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)

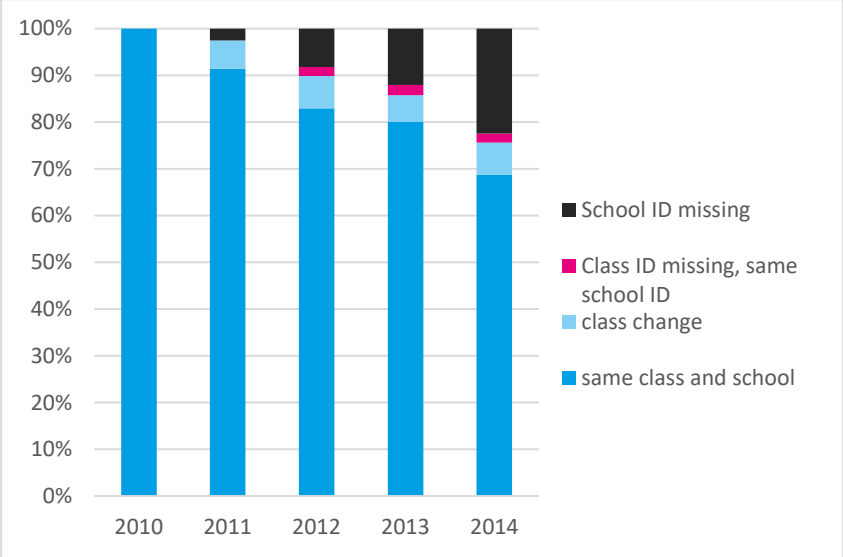

Figure 14. Overall mobility of Comprehensive schools students sampled in 2010. Authors' own elaboration of NEPS SC3 data (doi:10.5157/NEPS:SC3:7.0.1)