



NEPS WORKING PAPERS

Michael Bayer, Frank Goßmann and Daniel Bela

NEPS TECHNICAL REPORT:  
GENERATED SCHOOL TYPE  
VARIABLE T723080\_G1 IN STARTING  
COHORTS 3 AND 4

NEPS Working Paper No. 46  
Bamberg, July 2014

**Working Papers of the German National Educational Panel Study (NEPS)**

at the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).

The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at

<https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers>

**Editorial Board:**

Jutta Allmendinger, WZB Berlin

Cordula Artelt, University of Bamberg

Jürgen Baumert, MPIB Berlin

Hans-Peter Blossfeld, EUI Florence

Wilfried Bos, University of Dortmund

Edith Braun, DZHW Hannover

Claus H. Carstensen, University of Bamberg

Henriette Engelhardt-Wölfler, University of Bamberg

Frank Kalter, University of Mannheim

Corinna Kleinert, IAB Nürnberg

Eckhard Klieme, DIPF Frankfurt

Cornelia Kristen, University of Bamberg

Wolfgang Ludwig-Mayerhofer, University of Siegen

Thomas Martens, DIPF Frankfurt

Manfred Prenzel, TU Munich

Susanne Rässler, University of Bamberg

Marc Rittberger, DIPF Frankfurt

Hans-Günther Roßbach, LifBi

Hildegard Schaeper, DZHW Hannover

Thorsten Schneider, University of Leipzig

Heike Solga, WZB Berlin

Petra Stanat, IQB Berlin

Volker Stocké, University of Kassel

Olaf Struck, University of Bamberg

Ulrich Trautwein, University of Tübingen

Jutta von Maurice, LifBi

Sabine Weinert, University of Bamberg

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

# **NEPS Technical Report: Generated School Type Variable t723080\_g1 in NEPS-SUFs of Starting Cohorts 3 and 4**

*Michael Bayer, Frank Goßmann, Daniel Bela*

*Leibniz Institute for Educational Trajectories, Bamberg, Germany*

**E-mail address of lead author:**

michael.bayer@lifbi.de

**Bibliographic data:**

Bayer, M., Goßmann, F., & Bela, D. (2014). *NEPS Technical Report: Generated school type variable t723080\_g1 in Starting Cohorts 3 and 4* (NEPS Working Paper No. 46). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

## **NEPS Technical Report: Generated School Type Variable t723080\_g1 in Starting Cohorts 3 and 4**

### **Abstract**

The school type currently attended by students is an important piece of information for educational research. Because the German educational system is highly complicated, and students—especially in Grade 5 (SC3)—are too young to give exhaustive information about their school history, this information is collected by interviews with parents. This leads to two problems: (1) As episodic data about school history are collected, there is no variable providing cross-sectional information about the attended school type. (2) Because not all parents take part in our interviews, we have to deal with a large amount of unit nonresponse.

In this report we describe one possible strategy for solving these problems. This strategy consists of multiple steps in which (1) the current school episode of students with parent information is identified, (2) the true school type information is assigned at class level based on data from the parent interview, and (3) the class-specific school type information thus obtained is used to replace missing data of those students whose parents did not participate in the parent interview.

### **Keywords**

generated variable, currently attended school type, class context

## 1. Introduction

In analyzing NEPS data of starting cohorts (SC) that were sampled in school contexts (SC2-SC4), the school type variable is of special interest. Germany has a very complex school system, which is determined by the Federal States' sovereignty over education policy. This makes a detailed variable on school types a highly significant piece of information.

Feedback from the scientific community has shown that variables containing information about the school-specific sample structure ("stratum" or "school type") are often used to analyze the distribution of students across schools. As this information remains constant over time, this procedure is not recommendable for the analysis of later waves. Additionally, this sample information about schools is not much differentiated either, so that the complexity of the German education system and the possibilities of school trajectories cannot be reflected in an appropriate way.

Until now, the school type currently attended by students has not been available as cross-sectional information. Users have had to extract this information from the students' school history data collected by the parent interview. This school history contains all school-related episodes since the student's entrance to elementary school. In this spell structure, the current school episode can be reconstructed using time-related information. Because of this complicated measurement structure and the young age of the students, at least in the early waves of SC3, this particular information is being collected by the parent interview.

Data released for SC3 ([doi:10.5157/NEPS:SC3:2.0.0](https://doi.org/10.5157/NEPS:SC3:2.0.0)) as well as SC4 ([doi:10.5157/NEPS:SC4:4.0.0](https://doi.org/10.5157/NEPS:SC4:4.0.0)) contain a large amount of unit nonresponse of parents. Hence, for all students without a corresponding parent interview information about school history (and, respectively, about the type of school currently attended) is absent. In this report we present our procedure for generating a variable to bypass this problem by using students' context information. The basic idea behind the applied procedure is to use indirect information—in this case information gathered from the parents of the target persons' classmates.

Additionally, the procedure can be used as a validity-check routine, as parents are not always exactly sure about the type of school which their children are attending. To make transparent how we have generated our variable, we present all the relevant steps and decisions that we have made in this process and which users are unable to see in the final variable.

## 2. Measurement of the School History in NEPS Starting Cohorts 3 and 4

In order to obtain detailed information about the educational careers of students, the entire school history is measured. This information is collected by computer-assisted telephone interviews (CATI) conducted with one of the students' parents. Because of the complexity of the German education system, two different variables must be combined to reconstruct the appropriate school type information from the data: type of school and school track. Tables 1 and 2 show the corresponding questions and response categories. Parents whose children attend a "verbundene Haupt- und Realschule" [type of school in Berlin, Hesse, Mecklenburg-Western Pomerania and Lower Saxony offering basic and intermediate secondary education]

or a comprehensive school (“Gesamtschule”) are asked what school track their child attends there.

Table 1

*Question on Type of School*

---

**What school did <Name of the target child> attend there?**

---

Elementary school

Orientation stage [first two years of secondary education]

Hauptschule (school for basic secondary education)

Realschule (intermediate secondary school)

Verbundene Haupt- und Realschule (type of school in Berlin, Hesse, Mecklenburg-Western Pomerania and Lower Saxony offering basic and intermediate secondary education)

Comprehensive school (Gesamtschule)

Waldorf school (Rudolf Steiner school)

Gymnasium (type of school leading to upper secondary education and Abitur)

Special needs school/Remedial school

Other school

---

Table 2

*Question on Branch of School*

---

**What branch did <Name of the target child> attend there?**

---

Hauptschule branch

Realschule branch

Gymnasium branch

(So far) no division into school branches

---

Because the whole school history is measured, data collection yields longitudinal data. In the data set, each reported school episode is represented by one spell. For our specific purposes however, longitudinal school type information must be converted into cross-sectional information.

### 3. From Longitudinal to Cross-Sectional Information

To extract the information about the currently attended school type, the continuous school episode has to be identified first. Because it is school types of the general education system that are measured, only one continuous (“censored”) school episode per student can possibly remain. However, there are few students whose parents report multiple continuous school episodes in the data. To solve this problem, we have defined two criteria to identify

the “actual” continuous episode. We assume that this episode fulfills the following two criteria:

1. The smallest time gap relative to the last discontinuous episode<sup>1</sup>: This criterion is based on our expectation that school history, at least regarding the general educational system, is characterized as a continuing, uninterrupted process.
2. The comparatively higher school type<sup>2</sup>: This criterion is based on the fact that, in most empirical cases, one of several continuous episodes is an elementary school episode, despite the fact that students should have proceeded to secondary school already.

Figure 1 shows the corresponding Stata<sup>®</sup> syntax for these steps.

```

generate `last'=p723110_g1==1
preserve
keep if `last'==1
capture : isid ID_t
restore
if (_rc!=0) {
    generate `start' = ym(p72301y_g1, p72301m_g1)
    generate `end' = ym(p72302y_g1, p72302m_g1)
    drop if missing(`start') | missing(`end')
    gsort ID_t -`last' p72302y_g1 p72302m_g1
    bysort ID_t : generate `lastend'=`end'[_N]
    format %tm `lastend' `end' `start'
    generate `diff'=abs(`start'-`lastend')
    replace `diff'=. if `last'!=1
    bysort ID_t : egen `mindiff'=min(`diff')
    replace `last'=0 if (///
        (`last'==1 & `mindiff'!=`diff') ///
    )
    preserve
    keep if `last'==1
    capture : isid ID_t
    restore
    if (_rc!=0) {
        generate `schooltype'=.
        replace `schooltype'=1 if p723080==1
        replace `schooltype'=2 if p723080==2
        replace `schooltype'=3 if p723080==4
        replace `schooltype'=4 if p723080==5
        replace `schooltype'=5 if p723080==6
        replace `schooltype'=6 if p723080==8
        bysort ID_t (`last' `schooltype') : generate `maxschool'=`schooltype'[_N] if `last'==1
        replace `last'=0 if (`last'==1 & `schooltype'!=`maxschool')
        drop `schooltype' `maxschool'
    }
    sort ID_t spell
}
keep if `last'==1

```

Figure 1. Syntax for extracting the actual current school episode.

<sup>1</sup> Episodes that have missing values in either start- or end-date variable are disregarded.

<sup>2</sup> Elementary school < Orientation stage < Hauptschule < Realschule < Verbundene Haupt- und Realschule < Gymnasium.

#### 4. Process of Generating Class-Specific Information

After identifying the current school type of all students with the aid of observed parent interviews, we analyze this information in the context of the respective school classes. We can ascertain the class context for each student as well as their classmates. In contrast to our assumption that parents know what type of school their children are currently attending or were attending in the past (and, thus, current school type should be reported consistently by parents within each class), school type information within class contexts is oftentimes heterogeneous. To deal with this problem, we decided to use the statistical mode as an approximation to the actual school type attended. Therefore, we assume that the majority of parents' answers about the school context are identical to the true school context. By doing so, we further assume that any cases of children from the same class attending two different school types or tracks can be realistically ruled out. The heterogeneity of answers manifests on two levels:

*First* on the level of school type and, *second*, on the level of tracks (for schools with multiple tracks). As a consequence, there is not only one mode that is available. We use two strategies to solve these problems.

1. Parents sometimes misreport their children's current school type by reporting a school type which does not exist in their Federal State. To adjust this kind of structural impossibility, we use the corresponding information about the education system at state level to make sure that "false" data, in this sense, cannot affect the identification of the class-specific mode. This step is documented in Figure 2.

```
replace school = . if ///
(p723080 == 2 & linlist(bula, -, -97, -98, 1, 4, 5, 7, 8, 9, 13, 14, 15, 16)) | ///
(p723080 == 4 & linlist(bula, -, -97, -98, 1, 2, 3, 5, 6, 7, 8, 9, 11)) | ///
(p723080 == 5 & linlist(bula, -, -97, -98, 1, 2, 3, 5, 6, 7, 8, 9, 11)) | ///
(p723080 == 6 & linlist(bula, -, -97, -98, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16)) | ///
(p723080 == 10 & linlist(bula, -, -97, -98, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 15, 16))
```

Figure 2. Syntax for deleting structurally implausible information.

2. In schools with multiple tracks, we use a two-tier procedure to identify the applicable track (see Table 2) in case of multiple modes. At first we dichotomize information about the attended track into "integrated" (no division into school branches) and "non-integrated" school (Hauptschule branch, Realschule branch and Gymnasium branch), as we assume that this information, which is less detailed, is easier to recall for interviewees and thus is more valid than the more differentiated track information. Afterwards, we identify the statistical mode of the dichotomized school track information. If the statistical mode signifies that the school is a "non-integrated school", we carry out a second step to determine the correct track by using the statistical mode among Hauptschule branch, Realschule branch, and Gymnasium branch.

#### 5. Process of generating a differentiated school type variable

Using class-specific information about school type and school track, differentiated school type information can be generated (see Table 3). Details about school track are used to fur-



ther distinguish between school types offering basic and intermediate secondary education as well as comprehensive schools (Gesamtschulen).

As information required for generating the cross-sectional school type variable is only measured as part of the parent interview, missing data occur if students' parents do not participate in the interview. Hence, in a final step, missing data are replaced by the cross-sectional school type information of the students' classmates (see Figure 3) so that the missing rate is reduced remarkably (see Table 3). The remaining missing values occur due to two reasons: (1) None of the parents in the relevant class has participated in the parent interview or (2) there are multiple statistical modes concerning the variable regarding school type (see Table 1). If there are multiple statistical modes concerning school track variable (see Table 2) and only one mode concerning the type of school variable, we decided to keep this information by introducing Categories 8 ("School with several tracks: indeterminable") and 15 ("Comprehensive school: indeterminable") so that the user can decide whether or not to use it.

```

by sort ID_cc wave : generate `N' = _N if wave==1 & ID_cc>0 & !missing(ID_cc)
summarize `N', meanonly
forvalues i = 1 / `r(max)' {
    by sort ID_cc wave : replace `target_variable' = `target_variable'[`i'] if (missing(`target_variable') & !missing(`target_variable'[`i']))
    & wave==1 & ID_cc>0 & !missing(ID_cc)
}

```

Figure 3. Syntax for missing data replacement by class-specific information.

Table 3

*Comparison of Cross-Sectional School Type Information With and Without Class-Specific Information*

	cross-sectional school type information (generated variable) - t723080_g1	Information from parent SC4 W1 (in percent)	Class specific information SC4 W1 (in percent)
Value	Value Label		
1	Elementary school	0.83	0.08
2	Orientation stage	0.24	0.03
3	Hauptschule (school for basic secondary education)	16.47	22.16
4	Realschule (intermediate secondary school)	21.55	19.94
5	School with several educational programs: no separation into school branches (yet)	0.68	0.51
6	School with several educational programs: school branch of the Hauptschule	1.58	1.02
7	School with several educational programs: school branch of the Realschule	3.10	3.89
8	School with several educational programs: unclear	0.12	0.88
9	Gymnasium	38.14	32.72
10	Special school	5.31	6.61

11	Comprehensive school: no separation into school branches (yet)	3.16	3.19
12	Comprehensive school: school branch of the Hauptschule	1.64	1.14
13	Comprehensive school: school branch of the Realschule	2.91	3.73
14	Comprehensive school: school branch of the Gymnasium	2.40	1.94
15	Comprehensive school: unclear	0.23	1.03
16	Waldorf school	0.96	0.82
17	Other school	0.67	0.32
Total		100	100
Missing		44.30	3.93

---

## 6. Outlook

The generated variable about the current school type attended will be incorporated as a regular component of future releases of the NEPS Scientific Use Files. Users should be able to use this information to describe and analyze actual and differentiated distributions.

For each wave of parent interviews we will implement the cross-sectional information in the “CohortProfile” data set.

Because school trajectories are manifold, not all phenomena can be covered in this frame. For example, students change schools and are tracked individually. Regarding the school type attended by these students, only little information is collected. Our generated variable (t723080\_g1) is therefore useful for those students who stay within their specific institutional context as part of their educational trajectory—that is, it cannot be applied to individually tracked students.