

**CILS4NEPS**  
**A Harmonised Dataset Based on CILS4EU and**  
**NEPS SC4**

**Technical Report**

**CILS4EU Waves 1–3; NEPS SC4 Waves 1–6**

**2023**

Version: 1.0

Year: 2023

Citation: Dollmann, J., Arnold, L., Horr, A., Kerzner, V., Schmidt, R., Soiné, H., Weber, F. & Weißmann, M. (2023). Technical Report: A Harmonised Dataset Based on CILS4EU and NEPS SC4 (CILS4NEPS), Version 1.0. Mannheim: Mannheim University.

We are very thankful to Dr. Verena Ortmanns and Dr. Ranjit Singh from GESIS for their valuable insights and help in the creation of the harmonised CILS4NEPS dataset.

Furthermore, we thank our student assistants Denise Roth, Alena Nafe, Leoni Kotwan, and Dominik Keller for their excellent work. Especially Dominik Keller and Leoni Kotwan have contributed significantly to the construction and documentation of the CILS4NEPS dataset.

# Content

<b>1</b>	<b>Introduction.....</b>	<b>1</b>
<b>2</b>	<b>Sample.....</b>	<b>1</b>
2.1	Description of the CILS4EU and NEPS SC4 Target Population and Sample.....	1
2.2	Comparability of the Target Populations and Samples .....	3
2.3	Combined Sample.....	3
<b>3</b>	<b>Harmonisation Process and Methods .....</b>	<b>7</b>
3.1	Ex-Post Harmonisation .....	7
3.2	Preparation of the CILS4EU and NEPS SC4 Datasets .....	7
3.3	Steps of Ex-Post Harmonisation .....	8
3.3.1	Identification of Harmonisable Variables .....	8
3.3.2	Definition of Target Variables.....	9
3.3.3	Harmonisation Strategies .....	12
3.3.3.1	<i>Matching – Observed Constructs</i> .....	12
3.3.3.2	<i>Equating – Latent Constructs</i> .....	12
3.4	Structure of the Harmonised Dataset .....	17
3.4.1	Identifier Variable .....	18
3.4.2	Wave Indicators.....	18
3.4.3	Dealing with Duplicate Cases in the Harmonised Dataset.....	22
3.5	Weighting.....	23
3.6	Information on Data Use.....	24
<b>4</b>	<b>Special Harmonised Variables.....</b>	<b>24</b>
4.1	Generational Status and Ethnic Origin .....	24
4.1.1	General remarks on the collection of information in CILS4EU and NEPS SC4 .....	25
4.1.2	Coding Strategies: Generational Status .....	26
4.1.3	Coding Strategies: Ethnic Origin .....	33
4.2	Tracking Variables.....	33

<b>5</b>	<b>Construction of Filter Variables.....</b>	<b>34</b>
5.1	Balanced Panel.....	34
5.2	Selection of Items Classified as Unproblematic versus More Complicated.....	34
5.3	Linear Equating: Differences in Year-Overlap.....	35
<b>6</b>	<b>Missing Values.....</b>	<b>35</b>
<b>7</b>	<b>Appendix.....</b>	<b>38</b>
7.1	A – Documentation of Changes in the CILS4EU Data for the Converting from a Wide to a Long Data Format Converting .....	38
7.2	B – Documentation of Divergent Answers between Student and School Leaver Variables in Waves Three and Five .....	46
7.3	C – Documentation Year-Overlap Linear Equating .....	47
<b>8</b>	<b>References.....</b>	<b>49</b>

## List of Tables and Figures

<b>Table 1.</b> Sample Sizes.....	4
<b>Table 2.</b> Composition of the Individual Sample per Wave .....	5
<b>Table 3.</b> Composition of the Balanced Panel .....	6
<b>Table 4.</b> Overview of Dataset-Specific Classification Approaches for Generational Status .	29
<b>Table 5.</b> Cross-Tabulation of the Variable Generational Status According to the CILS4EU Approach (Rows) and According to the NEPS Approach (Column) Based on CILS4EU.....	32
<b>Table 6.</b> Missing Codes in the Harmonised Dataset.....	37
<b>Figure 1.</b> Example of an Unproblematic Item – ‘Day of Birth, Month’ .....	10
<b>Figure 2.</b> Example of a More Complicated Item – ‘Self-Efficacy – Do Well at School’ .....	11
<b>Figure 3.</b> Example of a Problematic Item – ‘Gender Roles – Child Care’ .....	12
<b>Figure 4.</b> Input Table .....	15
<b>Figure 5.</b> Recoding Table .....	15
<b>Figure 6.</b> Graphic Representation of the Linear Transformation and Recoding Values .....	16
<b>Figure 7.</b> Answer Scale of an Equated Harmonised Item .....	17
<b>Figure 8.</b> Excerpt from the Harmonised Dataset .....	18
<b>Figure 9.</b> Wave Structure in CILS4EU: Wave 1–3 .....	19
<b>Figure 10.</b> Wave Structure in NEPS SC4: Wave 1–6 (LifBi, 2021).....	20
<b>Figure 11.</b> Cross-Tabulation of the Two Harmonised Wave Indicators ‘H_wave2’ and ‘H_wave’ .....	21
<b>Figure 12.</b> Missing Codes in CILS4EU (CILS4EU, 2016).....	36
<b>Figure 13.</b> Missing Codes in NEPS SC4 (Skopek, Pink and Bela, 2013).....	36

# 1 Introduction

The aim of the harmonised dataset is to tap additional research potential and enable more differentiated analyses by combining the two data sources Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU; Kalter et al., 2016) and National Educational Panel Study: Starting Cohort 4 – 9<sup>th</sup> Grade” (in the following NEPS SC4; Blossfeld, Rossbach and Maurice, 2011). Combining both sources is a useful enrichment for national analyses, as it can increase the sample sizes of certain groups (e.g. ethnic or social groups) as well as of certain events (e.g. transitions to certain forms of school or education). Furthermore, the combination of the two datasets makes it possible to use NEPS SC4 for comparisons of school and labour market trajectories of young people in Germany with those in England, the Netherlands or Sweden (the three countries that participated in CILS4EU alongside Germany).

In the following, we briefly describe the CILS4EU and NEPS SC4 datasets, their target populations, and their comparability (Sections 2.1 and 2.2), before providing information on the composition of the harmonised dataset (Section 2.3), the harmonisation process (Section 3), special harmonised variables (Section 4), filter variables (Section 5), and the harmonised missing scheme (Section 6). We also provide documentation material accompanying the harmonised dataset (Section 7). For more detailed information about the individual CILS4EU and NEPS SC4 datasets, please refer to the respective technical reports and codebooks ([www.cils4.eu](http://www.cils4.eu); [www.neps-data.de](http://www.neps-data.de)).

## 2 Sample

### 2.1 Description of the CILS4EU and NEPS SC4 Target Population and Sample

#### CILS4EU

CILS4EU is an international longitudinal study investigating the structural, cultural, social, and emotional integration of young people with and without a migration background in Germany, England, the Netherlands, and Sweden. The study design is based on that of NEPS SC4. In addition to explicit stratification with an oversampling of migrant-rich schools, implicit stratification was performed in the individual countries – in Germany by federal state and type of school to account for these characteristics proportionally to the population.

A total of approximately 18,700 young people (aged 14–15 years) were surveyed in wave 1 (5,000 in Germany), about half of whom have a migration background. Within the period of international funding, two further survey waves were conducted from 2011 to 2013. The second

wave took place in the school context, while participants in the third wave were surveyed outside the school context – online, by post or by telephone. After the third wave, the German part of CILS4EU was integrated in the long-term programme of the Deutsche Forschungsgemeinschaft (DFG's) (continuing the data collection in Germany only). In 2016, during the sixth survey wave, a refreshment sample was drawn, in which adolescents and young adults with a migration background were also disproportionately represented. Currently, data from eight waves (plus one wave on the COVID-19 pandemic) are available; data for the ninth wave are currently being collected (June 2022).

#### NEPS SC4

NEPS collects longitudinal data on competence developments and educational returns in formal, non-formal, and informal contexts. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research, which was funded by the Bundesministerium für Bildung und Forschung (BMBF). Since 2014, NEPS has been continued by the Leibniz Institute for Educational Pathways (LifBi) in cooperation with a Germany-wide network.

The SC4 sub-study examines, starting from grade 9, pathways into and through upper secondary education as well as transitions to the vocational education system, higher education, and the labour market. The target population of the survey was the student body at regular schools and special schools who attended grade 9 in the school year 2010/11. For this purpose, a stratified cluster sample was drawn from regular schools and a sample of adolescents was drawn from special schools.

At the time of the first survey, adolescents were 14–15 years old. In total, survey data are available for approximately 15,500 adolescents, of whom about 37 percent have a migration background.<sup>1</sup> Young people who continued to attend the surveyed school were interviewed again in the school context in the following waves. School-leavers were followed up outside the school context (mainly via computer-assisted telephone interviewing, CATI).

The main survey at the schools was conducted by IEA-DPC through paper and pencil interviewing (PAPI), the CATI and computer-assisted web-interviewing (CAWI) surveys in the

---

<sup>1</sup> We define respondents as having a migration background up to the 3.5<sup>th</sup> immigrant generation. For a detailed overview of the construction of respondents' ethnic origin and generational status in the harmonised dataset, please refer to Section 4.1.

individual fields were conducted by infas (Institute for Applied Social Sciences). Currently, data from eleven waves (plus one wave on the COVID-19 pandemic) are available for SC4.

## **2.2 Comparability of the Target Populations and Samples**

Combining the two datasets is possible because CILS4EU and NEPS SC4 refer to the same target population (adolescents aged 14–15 years and older) – or in the case of the German part of CILS4EU – even to the same population (adolescents in grade 9 in the school year 2010/11 in Germany). The sampling approach at the school level in CILS4EU is very similar to that in NEPS SC4, although schools with high proportions of migrants were oversampled in CILS4EU. In CILS4EU, the international sampling as well as the national sampling and the fieldwork in the drawn schools in Germany were carried out by the IEA-DPC (also responsible for PISA, TIMSS, etc.), which was also responsible for the sampling design and large parts of the fieldwork in NEPS SC4. This additionally ensures comparability and a meaningful combination of the two data sources. Importantly, schools sampled in NEPS SC4 were originally excluded from the CILS4EU sample in Germany. This means that the same students do not appear twice in the harmonised dataset.

## **2.3 Combined Sample**

The combined sample of the harmonised dataset includes participants from CILS4EU waves 1–3 (youth main dataset) and NEPS SC4 waves 1–6 (pTarget und pTargetCATI datasets). This difference in included waves from the two source datasets results from a different frequency of data collection: NEPS SC4 collected data twice a year, with different waves for students and school-leavers, while data collection in the first three waves of CILS4EU was performed on a yearly basis for all respondents combined. Please refer to Section 3.4.2 for further information on the harmonisation of the CILS4EU and NEPS SC4 wave indicators.

The harmonised dataset comprises a total of 34,293 respondents in the first wave. Table 1 shows the sample sizes of both the source and the combined dataset(s) at the school, class, and individual level. Overall, 1,128 schools are included in the harmonised dataset, with 480 belonging to the CILS4EU and 648 to the NEPS SC4 sample. Across these schools, 2,147 school classes are included in the harmonised data. Of the 34,293 respondents, 21,103 (61.54 percent) participated in all three waves of the respective survey. These constitute the balanced panel (see last column of Table 1). Table 2 provides further detail on the composition of the individual sample per wave. In wave 1, 17,277 respondents are male (50.38 percent) and 17,009 are female (49.60 percent). Furthermore, 12,993 have a migration background (37.89 percent). An overview of the composition of the balanced panel is given in Table 3.



**Table 1. Sample Sizes**

	School level			Class level			Individual level			
	Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3	Balanced panel
Germany (combined)	792 (70.21 %)	791 (70.94 %)	788 (70.74 %)	1,466 (68.28 %)	1,197 (64.46 %)	860 (56.58 %)	20,590 (60.04 %)	18,262 (61.29 %)	16,402 (68.00 %)	14,406 (68.27 %)
<i>CILS4EU</i> (% of Germany combined)	144 (18.18 %)	144 (18.20 %)	142 (18.02 %)	271 (18.49 %)	268 (22.39 %)	264 (30.70 %)	5,013 (24.35 %)	4,256 (23.31 %)	3,427 (20.89 %)	3,260 (22.63 %)
<i>NEPS SC4</i> (% of Germany combined)	648 (81.82 %)	647 (81.80 %)	646 (81.98 %)	1,195 (81.51 %)	929 (77.61 %)	596 (69.30 %)	15,577 (75.65 %)	14,006 (76.69 %)	12,975 (79.11 %)	11,146 (77.37 %)
England <i>CILS4EU</i>	107 (9.49 %)	97 (8.70 %)	97 (8.71 %)	208 (9.69 %)	190 (10.23 %)	190 (12.50 %)	4,315 (12.58 %)	3,389 (11.37 %)	2,284 (9.47 %)	2,227 (10.55 %)
Netherlands <i>CILS4EU</i>	100 (8.87 %)	100 (8.97 %)	100 (8.98 %)	222 (10.34 %)	222 (11.95 %)	219 (14.41 %)	4,363 (12.72 %)	3,614 (12.13 %)	2,667 (11.06 %)	2,246 (10.64 %)
Sweden <i>CILS4EU</i>	129 (11.44 %)	127 (11.39 %)	129 (11.58 %)	251 (11.69 %)	248 (13.35 %)	251 (16.51 %)	5,025 (14.65 %)	4,531 (15.21 %)	2,768 (11.48 %)	2,224 (10.54 %)
<b>Total (combined)</b>	<b>1,128</b> <b>(100 %)</b>	<b>1,115</b> <b>(100 %)</b>	<b>1,114</b> <b>(100 %)</b>	<b>2,147</b> <b>(100 %)</b>	<b>1,857</b> <b>(100 %)</b>	<b>1,520</b> <b>(100 %)</b>	<b>34,293</b> <b>(100 %)</b>	<b>29,796</b> <b>(100 %)</b>	<b>24,121</b> <b>(100 %)</b>	<b>21,103</b> <b>(100 %)</b>
<i>CILS4EU</i>	480 (42.55 %)	468 (41.97 %)	468 (42.01 %)	952 (44.34 %)	928 (49.97 %)	924 (60.79 %)	18,716 (54.58 %)	15,790 (52.99 %)	11,146 (46.21 %)	9,957 (47.18 %)
<i>NEPS SC4</i>	648 (57.45 %)	647 (58.03 %)	646 (57.99 %)	1,195 (55.66 %)	929 (50.03 %)	596 (39.21 %)	15,577 (45.42 %)	14,006 (47.01 %)	12,975 (53.79 %)	11,146 (52.82 %)

**Table 2.** Composition of the Individual Sample per Wave

	Wave 1				Wave 2				Wave 3			
	Male	Female	With imm. backg.	Without imm. backg.	Male	Female	With imm. backg.	Without imm. backg.	Male	Female	With imm. backg.	Without imm. backg.
Germany (combined)	10,441 (60.43 %)	10,147 (59.66 %)	7,074 (54.44 %)	13,506 (63.44 %)	9,188 (61.38 %)	9,073 (61.28 %)	5,945 (54.82 %)	12,169 (64.72 %)	8,128 (69.85 %)	8,271 (66.27 %)	5,220 (62.63 %)	11,021 (70.54 %)
<i>CILS4EU</i> (% of Germany combined)	2,571 (24.62 %)	2,442 (24.07 %)	2,500 (35.34 %)	2,513 (18.61 %)	2,140 (23.29 %)	2,115 (23.31 %)	2,066 (34.75 %)	2,190 (18.00 %)	1,621 (19.94 %)	1,803 (21.80 %)	1,639 (31.40 %)	1,788 (16.22 %)
<i>NEPS SC4</i> (% of Germany combined)	7,870 (75.38 %)	7,705 (75.93 %)	4,574 (64.66 %)	10,993 (81.39 %)	7,048 (76.71 %)	6,958 (76.69 %)	3,879 (65.25 %)	9,979 (82.00 %)	6,507 (80.06 %)	6,468 (78.20 %)	3,581 (68.60 %)	9,233 (83.78 %)
England <i>CILS4EU</i>	2,211 (12.80 %)	2,102 (12.36 %)	2,013 (15.49 %)	2,301 (10.81 %)	1,764 (11.78 %)	1,623 (10.96 %)	1,632 (15.05 %)	1,756 (9.34 %)	1,117 (9.60 %)	1,167 (9.35 %)	1,075 (12.90 %)	1,209 (7.74 %)
Netherlands <i>CILS4EU</i>	2,144 (12.41 %)	2,216 (13.03 %)	1,469 (11.31 %)	2,894 (13.59 %)	1,761 (11.76 %)	1,851 (12.50 %)	1,145 (10.56 %)	2,469 (13.13 %)	1,203 (10.34 %)	1,464 (11.73 %)	729 (8.75 %)	1,937 (12.40 %)
Sweden <i>CILS4EU</i>	2,481 (14.36 %)	2,544 (14.96 %)	2,437 (18.76 %)	2,588 (12.16 %)	2,256 (15.07 %)	2,258 (15.25 %)	2,122 (19.57 %)	2,409 (12.81 %)	1,189 (10.22 %)	1,579 (12.65 %)	1,311 (15.73 %)	1,457 (9.33 %)
<b>Total (combined)</b>	<b>17,277</b> <b>(100 %)</b>	<b>17,009</b> <b>(100 %)</b>	<b>12,993</b> <b>(100 %)</b>	<b>21,289</b> <b>(100 %)</b>	<b>14,969</b> <b>(100 %)</b>	<b>14,805</b> <b>(100 %)</b>	<b>10,844</b> <b>(100 %)</b>	<b>18,803</b> <b>(100 %)</b>	<b>11,637</b> <b>(100 %)</b>	<b>12,481</b> <b>(100 %)</b>	<b>8,335</b> <b>(100 %)</b>	<b>15,624</b> <b>(100 %)</b>
<i>CILS4EU</i>	9,407 (54.45 %)	9,304 (54.70 %)	8,419 (64.80 %)	10,296 (48.36 %)	7,921 (52.92 %)	7,847 (53.00 %)	6,965 (64.23 %)	8,824 (46.93 %)	5,130 (44.08 %)	6,013 (48.18 %)	4,754 (57.04 %)	6,391 (40.91 %)
<i>NEPS SC4</i>	7,870 (45.55 %)	7,705 (45.30 %)	4,574 (35.20 %)	10,993 (51.64 %)	7,048 (47.08 %)	6,958 (47.00 %)	3,879 (35.77 %)	9,979 (53.07 %)	6,507 (55.92 %)	6,468 (51.82 %)	3,581 (42.96 %)	9,233 (59.09 %)

*Note.* Individuals are defined as having an immigrant background if at least one maternal and one paternal grandparent is born abroad. For detailed definitions of the immigrant generations and the differences between CILS4EU and NEPS, see Section 4.1. Missing values on the immigrant background variable are replaced with the highest value; missing values on the sex variable are replaced with the first non-missing value.

**Table 3.** Composition of the Balanced Panel

	Male	Female	With imm. backg.	Without imm. backg.	Total
Germany (combined)	7,135 (69.90 %)	7,271 (66.74 %)	4,590 (63.10 %)	9,811 (70.97 %)	14,406 (68.27 %)
<i>CILS4EU</i> (% of Germany combined)	1,554 (21.78 %)	1,706 (23.46 %)	1,555 (33.88 %)	1,705 (17.38 %)	3,260 (22.63 %)
<i>NEPS SC4</i> (% of Germany combined)	5,581 (78.22 %)	5,565 (76.54 %)	3,035 (66.12 %)	8,106 (82.62 %)	11,146 (77.37 %)
England <i>CILS4EU</i>	1,087 (10.65 %)	1,140 (10.46 %)	1,063 (14.61 %)	1,164 (8.42 %)	2,227 (10.55 %)
Netherlands <i>CILS4EU</i>	1,014 (9.93 %)	1,232 (11.31 %)	591 (8.12 %)	1,655 (11.97 %)	2,246 (10.64 %)
Sweden <i>CILS4EU</i>	972 (9.52 %)	1,252 (11.49 %)	1,030 (14.16 %)	1,194 (8.64 %)	2,224 (10.54 %)
<b>Total (combined)</b>	<b>10,208</b> <b>(100 %)</b>	<b>10,895</b> <b>(100 %)</b>	<b>7,274</b> <b>(100 %)</b>	<b>13,824</b> <b>(100 %)</b>	<b>21,103</b> <b>(100 %)</b>
<i>CILS4EU</i>	4,627 (45.33 %)	5,330 (48.92 %)	4,239 (58.28 %)	5,718 (41.36 %)	9,957 (47.18 %)
<i>NEPS SC4</i>	5,581 (54.67 %)	5,565 (51.08 %)	3,035 (41.72 %)	8,106 (58.64 %)	11,146 (52.82 %)

*Note.* Individuals are defined as having an immigrant background if at least one maternal or paternal grandparent is born abroad.

For detailed definitions of immigrant generations and the differences between CILS4EU and NEPS, see Section 4.1. Missing values on the immigrant background variable are replaced with the highest value; missing values on the sex variable are replaced with the first non-missing value.

### **3 Harmonisation Process and Methods**

This section provides an overview of the process and methods used during the harmonisation procedure combining CILS4EU and NEPS SC4. We first offer information on the general process of ex-post harmonisation and describe which criteria we used to identify the variables in the CILS4EU and NEPS SC4 that could be harmonised (Section 3.1). Subsequently, we outline the changes we made to the individual datasets before harmonisation could take place (Section 3.2). We further outline the applied harmonisation methods (Section 3.3). In Section 3.4, 3.5, and 3.6, we provide information on the structure of the harmonised dataset and data use, respectively. Only data on students and school-leavers were used for harmonisation; no other information from additional questionnaires of CILS4EU or NEPS SC4 on other groups of persons, such as parents or teachers, were harmonised.

#### **3.1 Ex-Post Harmonisation**

In contrast to ex-ante harmonisation, in which surveys are designed in a way to be comparable before data collection, ex-post harmonisation refers to the harmonisation of already existing survey data into one integrated dataset (Granda, Wolf and Hadorn, 2010). The goal of ex-post harmonisation is to construct a combined dataset with harmonised variables that originate from different source datasets but build on a common definition of construction (Wolf et al., 2016). This process can be used for cross-national as well as national surveys. For the CILS4EU and NEPS SC4 harmonisation, ex-post harmonisation is the available strategy because both studies have been collecting data for certain periods of time already.

The main benefit of data harmonisation, in general, is that it opens up new research potential by, for instance, '[...] filling gaps in the data, increasing sample sizes, and improving the robustness and reproducibility of results' (Singh, 2020/2021). For the CILS4EU and NEPS SC4 data, harmonisation is especially valuable as it increases the sample size for certain groups (e.g. ethnic or social groups) and certain events (e.g. transitions to certain forms of school or education) and enables international comparisons for the NEPS SC4 dataset.

#### **3.2 Preparation of the CILS4EU and NEPS SC4 Datasets**

Before starting the harmonisation procedure, the CILS4EU and NEPS SC4 datasets had to be prepared to combine them. The CILS4EU data is provided in a wide data format, with every respondent corresponding to one row in the data. To be compatible with the NEPS SC4 data, which is provided in a long data format (multiple rows per respondent – one per wave), the first three waves of CILS4EU were combined and converted into a long data format as well. Due to

its wide format structure, every variable in the original CILS4EU dataset includes a prefix (e.g. ‘y1\_’) indicating the wave from which the respective variable originates. Even though the items are kept similar across the survey waves, there are small deviations in the answer categories. Therefore, when converting the CILS4EU data into a long data format, small changes in the labelling of the answer categories for certain variables were necessary. For instance, for the religion variable ‘y3\_rell’ in wave three, the answer category ‘Christianity: other’ had to be combined under the answer category ‘Christianity’. We document these changes to the source datasets in Appendix A (Section 7.1) of this technical report. To indicate variables from the CILS4EU dataset, we added the prefix ‘CILS4EU\_’ to all those variables.

For the NEPS SC4 data, no changes in the structure of the data were necessary. We identified the necessary datasets: ‘CohortProfile’, ‘pTarget’, ‘pTargetCATI’, and ‘Weights’ (from Version 13) and merged them according to the merging matrix provided by the NEPS data centre. As with the CILS4EU dataset, we added the prefix ‘NEPS\_’ to all variables from NEPS SC4 before combining both datasets.

### **3.3 Steps of Ex-Post Harmonisation**

No established steps for ex-post data harmonisation exist, but experts suggest the following procedure (Granda, Wolf and Hadorn, 2010; Singh, 2021): 1) identification of the source datasets to be combined, 2) identification of similar questions in the source questionnaires with potential for harmonisation, 3) definition of target variables combining the source variables into harmonised variables, 4) definition of and decision on harmonisation strategies to create the target variable, and 5) mapping of routines applied during the data harmonisation to ensure replicability. We followed these suggestions in the CILS4EU and NEPS SC4 harmonisation process and outline below how each of these suggestions was implemented.

#### **3.3.1 Identification of Harmonisable Variables**

The identification of harmonisable variables in different datasets is not trivial since for ex-post harmonisation to be feasible it is crucial that variables or instruments in each dataset measure similar constructs (Singh, 2020/2021). Even though variables do not have to be measured in the exact same way, a certain amount of similarity is necessary (Singh, 2020/2021). Combining variables that do not measure a similar construct in the different source datasets (i.e. concept mismatch) may introduce serious bias in the harmonised dataset (Singh, 2020/2021).

In the CILS4EU and NEPS SC4 harmonisation process, harmonisable variables were identified and assessed in the following way: Based on the questionnaires of CILS4EU waves 1–3, a list of all variables included in these three waves was constructed. Subsequently, for each variable

on this list, trained student assistants checked whether the NEPS SC4 questionnaire included a variable measuring a similar construct. If this was the case, the matching NEPS SC4 variable was added to the list, including its name, question wording, response categories, the wave the question was asked, and to which respondent groups it was asked. The information was gathered in an Excel table, which we provide as documentation material (the Overview Table). To classify the similarity of each CILS4EU and NEPS SC4 variable, we then introduced a colour scheme: Variables with a green background are (almost) identical regarding their question wording and response categories. Variables with a yellow background capture similar constructs but deviate somewhat in their question wording and/or response categories. Variables with a red background exist only in the CILS4EU but not in the NEPS SC4 dataset. This overview table formed the basis for our further steps in the ex-post harmonisation of variables and provides documentation, transparency, and traceability.

### 3.3.2 Definition of Target Variables

To define the target variables that combine the source variables of both datasets into harmonised variables, a coding table was constructed (see documentation material). This table offers a precise overview of how each source variable was coded, also including information on the coding of missing values. Each Excel sheet in the coding table corresponds to a content area of the overview table. The coding table thus provides an exact working template for each target variable by illustrating how the respective CILS4EU and NEPS SC4 variables needed to be recoded – the starting point for the creation of each target item.

To assess the comparability of the source variables, we drew on relevant literature on ex-post harmonisation of survey data (described above; Wolf *et al.*, 2016; Granda, Wolf and Hadorn, 2010; Hoffmeyer-Zlotnik, 2008) as well as advice from Dr. Verena Ortmanns and Dr. Ranjit Singh from GESIS. This involved checking whether the source variables from CILS4EU and NEPS SC4 measure the same construct with regard to the question wording and whether this construct is observable/manifest (e.g. education level, number of persons in a household) or a latent construct (non-observable, e.g. attitudes). Next, the similarity of the response categories was examined – both in terms of content as well as in form and number of the response categories. This assessment determined the harmonisation strategy to be applied to the item later. Comparability between CILS4EU and NEPS SC4 was classified as either *unproblematic*, *more complicated* or *problematic* for each item, which is also indicated in the coding table.

*Unproblematic* items are based on variables that measure observable constructs in the source datasets. An example of such an item is the variable ‘day of birth, month’, which in both

CILS4EU and NEPS SC4 is asked with the question ‘When were you born?’ and coded with numbers ranging from 1 to 12 for the answer categories ‘1 – January’ to ‘12 – December’ (see Figure 1). Due to this agreement in construct similarity and identical or very similar response categories, a harmonisation procedure could be carried out for this target item in the form of a simple matching (and, if necessary, recoding) of the response categories.

Item: Day of birth, Month	Item: Day of Birth - Mont	CILS4EU	NEPS
2	Question number	2	241 : 2
H_dobm Day of birth, Month	Variable	y3_dobm Day of Birth, Month	t70004m Day of Birth, Month
CILS: y*_dobm	Data record		pTarget
NEPS: t70004m & t70000m			
	Filter		
	Filter_contentual		
	Introduction		
When were you born (Month)?	Question text	When were you born? Month	When were you born?
	Submit list		
	Instruction		
	Interviewer Instruction		
	<b>Answer categories:</b>		
1 January	January	1 January	1 January
2 February	February	2 February	2 February
3 March	March	3 March	3 March
4 April	April	4 April	4 April
5 May	May	5 May	5 May
6 June	June	6 June	6 June
7 July	July	7 July	7 July
8 August	August	8 August	8 August
9 September	September	9 September	9 September
10 October	October	10 October	10 October
11 November	November	11 November	11 November
12 December	December	12 December	12 December
-44 Interrupted interview	Missing 1	-44 Interrupted Interview	
-52 Implausible value removed	Missing 2		-52 Implausible value removed
-54 Missing by design	Missing 3		-54 Missing by design
-55 Other missing	Missing 4	-55 Other missing	
-66 Question not asked	Missing 5	-66 Question not asked	
-88 No answer	Missing 6	-88 No answer	
-90 Unspecific missing	Missing 7		-90 Unspecific missing
-95 Implausible value	Missing 8		-95 Implausible value
-98 Don't know	Missing 9		-98 Don't know
-99 Filtered	Missing 10		-99 Filtered
	Generated variable		
12	Number of response categories	12	12
	Data remark		

**Figure 1.** Example of an *Unproblematic* Item – ‘Day of Birth, Month’

*More complicated* target items measure latent constructs in both datasets and are similar in terms of construct measures and response categories. They can also be observed constructs that require more than simple matching of the answer categories (e.g. by lagging response). An example of a latent item is ‘self-efficacy at school’, recorded in CILS4EU by asking ‘How much do you agree or disagree with each of these statements? I am sure that I can do well at school.’, with answer categories ranging from ‘1 – strongly agree’ to ‘5 – strongly disagree’, and in the NEPS SC4 by asking ‘How are you doing in school? I’m good in most subjects.’, with answer categories from ‘4 – does completely apply’ to ‘1 – does not apply at all’ (see Figure 2). For these items, a simple assignment of response categories would have led to a bias

in the dataset and analyses. Instead, we applied linear equating as a harmonisation strategy, which we outline in Section 3.3.3.2. We recommend users of these items to validate them in their analyses (see Section 3.6).

Item: Self-efficacy - do well at school	Item: Self-efficacy I	CILS4EU	NEPS
54	Question number	16	740 : 55214i
H_sesch Self-efficacy - do well at school	Variable	y2_seff1 Self-efficacy I	t66002c Self-concept school: good in most school subjects
	Data record		pTarget
CILS: y*_seff1 NEPS: t66002c	Filter		
	Filter_contentual		
	Introduction		
Agreement/ disagreement: I am good at school.	Question text	How much do you agree or disagree with each of these statements? I am sure that I can do well at school.	[NCS] How are you doing in school? I'm good in most subjects.
	Submit list		
	Instruction		
	Interviewer Instruction		
	<b>Antwortkategorien:</b>		
Rational Strongly agree / Does completely apply numbers	Strongly agree / Does completely apply	1 Strongly agree	4 does completely apply
Rational . numbers	Agree / Does rather apply	2 Agree	3 does rather apply
Rational . numbers	Neither agree nor disagree	3 Neither agree nor disagree	
Rational . numbers	Disagree / Does rather not apply	4 Disagree	2 does rather not apply
Rational Strongly disagree / Does not apply at all numbers	Strongly disagree / Does not apply at all	5 Strongly disagree	1 does not apply at all
-44 Interrupted interview	Missing 1	-44 Interrupted interview	
-54 Missing by design	Missing 2		-54 Missing by design
-55 Other missing	Missing 3	-55 Other missing	
-66 Question not asked	Missing 4	-66 Question not asked	
-88 No answer	Missing 5	-88 No answer	
-90 Unspecific missing	Missing 6		-90 Unspecific missing
-95 Implausible value	Missing 7		-95 Implausible value
	Generated variable		
~	Number of response	5	4
	Data remark		

**Figure 2.** Example of a *More Complicated* Item – ‘Self-Efficacy – Do Well at School’

*Problematic* items are variables that are included in both source datasets but contain too many deviations (in their question wording and/or response categories) so that construct comparability is no longer given. These variables were not harmonised, as this would have introduced serious bias in the harmonised dataset. An example of a problematic item is ‘gender roles – child care’. In the CILS4EU questionnaire this item is asked as a single-barrel item ‘In a family, who should do the following? Take care of the children’, while in NEPS SC4 the item is asked with a double meaning ‘It’s the man’s job to earn money and the woman’s job to take care of the household and family.’ (see Figure 3). Due to this difference in measurement, answers for the source items could not be mapped meaningfully to exactly one/the same answer category in a harmonised item.



CILS4EU				NEPS		
Gender roles: Child care	grol1	In a family, who should do the following? Take care of the children	Mostly the man Mostly the woman Both about the same	t44613a	It's the man's job to earn money and the woman's job to take care of the household and family.	completely disagree rather disagree rather agree completely agree

**Figure 3.** Example of a *Problematic* Item – ‘Gender Roles – Child Care’

### 3.3.3 Harmonisation Strategies

To generate the harmonised variable in the data, we first decided on the harmonisation strategy to be applied. The decision on the strategy was related to the classification into *unproblematic* or *more complicated* items introduced above. While we applied simple matching of response categories for unproblematic items, we used linear equating for more complicated items. Both harmonisation strategies are outlined below. We used the statistics program Stata Version 17 (StataCorp., 2021) to create all target items. Do-files for the harmonisation are provided in the documentation material.

#### 3.3.3.1 Matching – Observed Constructs

To harmonise variables classified as unproblematic, we merged the answer categories of both source variables. Depending on the coding of the source variables, this sometimes required reverse coding of the response scale or combining several answer categories into one category. For example, for the harmonised variable ‘repetition of school year’, the CILS4EU variable distinguished ‘yes’ answers into several categories (‘yes, in primary school’, ‘yes, in secondary school’, ‘yes, in primary and secondary school’), while the NEPS SC4 item only differentiated between ‘yes’ and ‘no’. In such cases, the harmonised item consists of the highest common denominator between the two source variables – in this case, yes/no answer categories.

In other cases, multiple variables from CILS4EU or NEPS SC4 were combined into one harmonised variable. An example of this is ‘household members’ (e.g. mother, stepmother, adoptive mother), which in NEPS SC4 are assessed with a combined question and in CILS4EU with separate questions. Therefore, the harmonised item ‘household members – mother, stepmother, adoptive mother’ is based on three CILS4EU variables and one NEPS SC4 variable.

#### 3.3.3.2 Equating – Latent Constructs

As stated above, for variables that measure latent (non-observable) constructs and are thus classified as more complicated, a simple matching of answer categories would lead to bias.

Since respondents' true scores on the latent constructs are unobservable, two respondents with the same true score could classify themselves on different scores of the same answer scale for two source items that intend to measure the same latent construct but differ slightly in their question wording (Singh, 2020/2021; Singh, 2020; Kolen and Brennan, 2014). The same is possible vice versa, when the same latent construct is measured on different answer scales, for example with one scale ranging from 1 to 5 and the second from 1 to 4. In these cases, equating becomes necessary.

For equating to produce reliable results when used as a harmonisation strategy, three prerequisites need to be fulfilled. First, it is assumed that respondents have one true score on the latent item, regardless of the survey – even though the observed scores may differ between surveys. This assumption is also referred to as equity property (Singh, 2021). Second, and as outlined above, in both surveys the items to be harmonised should be measured in a similar way. Third, to avoid temporal and spatial (e.g. cultural) differences, it is crucial that the samples from each source dataset refer to the same population and that items that are equated were surveyed within the same time (Kolen and Brennan, 2014; Singh, 2020). As described above, we fulfil these pre-requisites in the CILS4EU and NEPS SC4 harmonisation because both surveys encompass the same target population (see Section 2.2) and we carefully assessed the comparability of the source items within the coding table and classification scheme. If source items were measured in different years in CILS4EU and NEPS SC4, we decided to also harmonise these items but indicate the deviance in survey year in a filter variable (see Section 5.3).

### *Linear Equating*

For data harmonisation, we used linear equating as one sub-form of equating. When applying linear equating, we assume that differences in the distribution of the observed scores are only due to differences between the measurement instruments. Therefore, we align the distribution of answers in the different surveys (Singh, 2020). An important assumption in this harmonisation strategy is that the distributions of both items approximately follow a normal distribution (i.e. only differing in mean and standard deviation; Singh, 2021). When linearly equating the source item to the target items, values of the source items are linearly transformed – meaning that the mean and standard deviation of the source and target item become equal (Singh, 2021; Kolen and Brennan, 2014). As described by Singh (2021), 'respondents now have very similar scores on the transformed source instrument and the target instrument depending

on their position along the normal distribution. Respondents with the same z-score have the same harmonised score but scaled to the format of the target scale.’ (p. 128).

The linear transformation is performed according to the following formula, in which variable X is transformed to the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of variable Y (Kolen and Brennan, 2014).

$$I_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[ \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \right]. \quad \text{slope} = \frac{\sigma(Y)}{\sigma(X)}, \quad \text{and} \quad \text{intercept} = \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X).$$

### *Application of Linear Equating in the Data Harmonisation*

We selected the CILS4EU variables as target items. This means that we adapted the scale of each NEPS SC4 item to that of the respective CILS4EU target item. The linear transformations were conducted in a Microsoft Excel spreadsheet, which we programmed to automatically apply the above-stated formula. This provided us with a table indicating how each answer category of the NEPS SC4 item needed to be recoded to match the target item. When entering the frequencies of both the source and the target item’s answer categories in the Excel sheet, we only counted the German respondents of the CILS4EU survey. This is because equating requires the same target populations to produce a reliable recoding table, which can subsequently be applied to the whole sample. This means that we not only equated the German part of the CILS4EU data with the NEPS SC4 data on that basis but also the CILS4EU data from England, the Netherlands, and Sweden.

We weighted the data during the linear equating process with the applicable individual weights (hous\_wgt for the CILS4EU; w\_t for NEPS SC4) to achieve a valid representation of the sample populations (this weighting was relevant only to obtain the correct frequency distribution of each source item but is not included in the harmonised item itself).<sup>2</sup>

Figure 4 shows the input table of the Excel sheet in which the respective weighted frequencies for the target (CILS4EU) and source items (NEPS SC4) were entered. Note that only German cases were included for the frequency distribution of the CILS4EU item, as explained above.

---

<sup>2</sup> We conducted robustness checks by using wave specific (cross-sectional) weights in NEPS SC4 (e.g. wt\_1, wt\_2) in the linear equating process. These robustness checks revealed largely similar results compared to the use of longitudinal weights for NEPS SC4 (i.e. w\_t). For CILS4EU cases, only respondents who participated in wave 1 were included in the weighted frequency distribution.

	A	B	C	I	E	F	G	H
1		<b>CILS</b>			<b>NEPS</b>			<b>Variable name</b>
2		<b>Value</b>	<b>Count</b>		<b>Count</b>	<b>Value</b>		H_spm
3		<b>0</b>	0		0	<b>0</b>		
4		<b>1</b>	598.67		3229.62	<b>1</b>		
5		<b>2</b>	1542.16		6017.36	<b>2</b>		
6		<b>3</b>	1975.42		2339.04	<b>3</b>		
7		<b>4</b>	728.23		1512.71	<b>4</b>		
8		<b>5</b>	159.62		0	<b>5</b>		
9		<b>6</b>	0		0	<b>6</b>		
10		<b>7</b>	0		0	<b>7</b>		
11		<b>8</b>	0		0	<b>8</b>		
12		<b>9</b>	0		0	<b>9</b>		
13		<b>10</b>	0		0	<b>10</b>		

**Figure 4.** Input Table

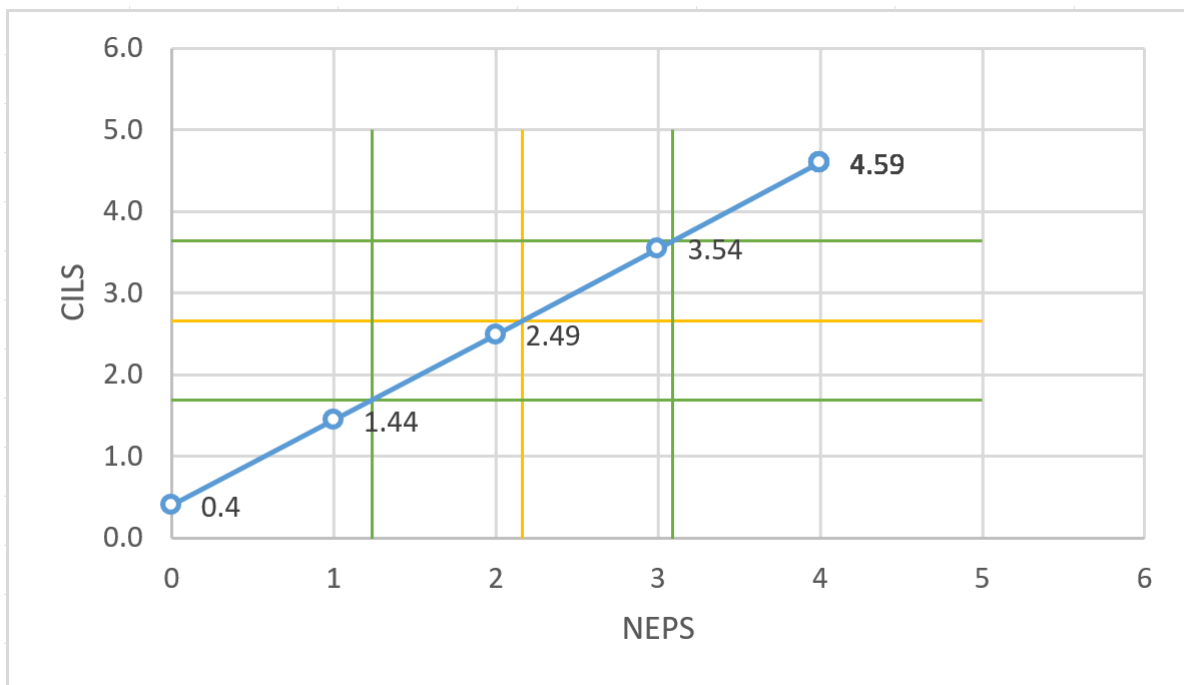
Based on this input table, the NEPS SC4 answer categories on the right-hand side were linearly transformed to align with the CILS4EU answer categories on the left. Figure 5 shows the recoding table (green numbers in the middle) resulting from this linear transformation.

	A	B	C	F	I	J
1		<b>H_spm</b>				
2		<b>CILS</b>			<b>NEPS</b>	
3		<b>Value</b>	<b>Count</b>	<b>MATCHED</b>	<b>Count</b>	<b>Value</b>
5		<b>0</b>	0	0.00	0	<b>0</b>
6		<b>1</b>	598.67	1.44	3229.62	<b>1</b>
7		<b>2</b>	1542.16	2.49	6017.36	<b>2</b>
8		<b>3</b>	1975.42	3.54	2339.04	<b>3</b>
9		<b>4</b>	728.23	4.59	1512.71	<b>4</b>
10		<b>5</b>	159.62	0.00	0	<b>5</b>
11		<b>6</b>	0	0.00	0	<b>6</b>
12		<b>7</b>	0	0.00	0	<b>7</b>
13		<b>8</b>	0	0.00	0	<b>8</b>
14		<b>9</b>	0	0.00	0	<b>9</b>
15		<b>10</b>	0	0.00	0	<b>10</b>
17		Total	5004.1		13098.73	Total

**Figure 5.** Recoding Table

In the example in Figure 5, the recoding table indicates that e.g. the value ‘1’ in the NEPS SC4 data needs to be recoded to ‘1.44’ for the harmonised item to match the answer scale of the CILS4EU item.

Figure 6 illustrates the equating procedure graphically. The x-axis represents the response categories of the NEPS SC4 dataset, and the y-axis represents the CILS4EU answer categories. The yellow line indicates the mean values of each item (CILS4EU and NEPS SC4). The green line represents exactly one standard deviation from the mean value of each item. The equating line (blue) is obtained by drawing a line through the intersections of the yellow and green lines. To obtain the numerical values stated in the recoding table above, we followed the numbers of the x-axis (‘1’) upwards to the blue line and read the value on the y-axis on the left (‘1.44’).



**Figure 6.** Graphic Representation of the Linear Transformation and Recoding Values

Based on the recoding table obtained in the Excel spreadsheet, the answer categories of the NEPS SC4 item were then recoded. The example in Figure 5 shows that all observations with the value 2 correspond to the value 2.49 in the recoding, all observations with the value 3 correspond to the value 3.54 in the recoding, and so on. This resulted in a scale for the harmonised item that preserves the CILS4EU answer scale but includes middle categories for the recoded NEPS SC4 values. We retained the labels for the answer categories from the CILS4EU data; the newly created values were not labelled (see

).

H\_spm — H 27: Subjective school performance, Math

		Freq.	Percent	Valid	Cum.
Valid	-95 Implausible value	37	0.03	0.04	0.04
	-90 Unspecific missing	236	0.21	0.28	0.33
	-88 No answer	148	0.13	0.18	0.50
	-77 Not applicable	28	0.03	0.03	0.54
	-66 Question not asked	850	0.77	1.02	1.55
	-55 Other missing	13	0.01	0.02	1.57
	-54 Missing by design	33642	30.56	40.18	41.75
	-44 Interrupted Interview	3	0.00	0.00	41.75
	1 Very well	6343	5.76	7.58	49.33
	1.3	3270	2.97	3.91	53.23
	2 Quite well	11676	10.61	13.95	67.18
	2.37	6167	5.60	7.37	74.55
	3 OK	10417	9.46	12.44	86.99
	3.44	4385	3.98	5.24	92.22
	4 Not that well	3687	3.35	4.40	96.63
	4.51	1482	1.35	1.77	98.40
	5 Not well at all	1341	1.22	1.60	100.00
	Total	83725	76.05	100.00	
Missing	.	26361	23.95		
Total		110086	100.00		

**Figure 7.** Answer Scale of an Equated Harmonised Item

If the target (CILS4EU) and source (NEPS SC4) items were originally coded in opposite directions, we conducted a correlation test to check whether the linear transformation produced correct results. This test had to result in the value of -1. After the linear transformation and recoding, we compared the NEPS SC4 item to the CILS4EU item – for which the mean values and standard deviations had to be exactly the same (apart from differences in rounding).

### 3.4 Structure of the Harmonised Dataset

Based on the combination of CILS4EU and NEPS SC4, the harmonised dataset is provided in a long data format. In this format, observations per respondent are distributed across several rows with a variable, in our case ‘H\_wave’ or ‘H\_wave\_2’, indicating the waves in which the variable was asked. The harmonised dataset includes the CILS4EU and the NEPS SC4 data as well as their harmonised variables. To provide a quick overview to which dataset the variables belong to, we included the label prefix ‘CILS4EU\_’ for all original CILS4EU variables and ‘NEPS\_’ for all original NEPS SC4 variables. Harmonised variables contain the prefix ‘H\_’ both in their variable name and in their variable label. We further included the variable ‘H\_dtset’, which indicates from which dataset the observations originate. See Figure 8 for an excerpt from the data structure.

Name	Label
H_dtset	H: Dataset Information
H_wave	H: Harmonized wave indicator 1-6
H_wave2	H: Harmonized wave indicator 1-3
H_ID	H: CILS4EU(8-digit) & NEPS_SC4(7-digit) ID
H_trackingC	H: Tracking variable: CILS4EU waves
H_trackingN	H: Tracking variable: NEPS SC4 waves
H_filter_balanced	H: Filter: Participation in all waves
H_sex	H 1: Sex Indicator
H_dobm	H 2: Day of birth, Month
H_doby	H 3: Day of birth, Year
H_hhm1	H 4: Household members: Biological mother, adoptive mothe...
H_hhm2	H 5: Household members: Biological father, adoptive father,...
H_hhm3	H 6: Household members: Stepmother
H_hhm4	H 7: Household members: Stepfather
H_hhm5	H 8: Household members: Siblings and/or Stepsiblings
H_hhm6	H 9: Household members: Grandparents
H_hhm7	H 10: Household members: Other family members
H_hhm8	H 11: Household members: Other persons
H_hhm9	H 12: Household size
H_edum	H 13: Mother's education (broad)
H_eduf	H 14: Father's education (broad)
H_empsm	H 15: Employment status mother
H_empsf	H 16: Employment status father
H_ocmisco	H 17: Occupation mother ISCO-08
H_ocfisco	H 18: Occupation father ISCO-08
H_counSC	H 19: Born in Survey Country
H_miga	H 20: Age at migration

**Figure 8.** Excerpt from the Harmonised Dataset

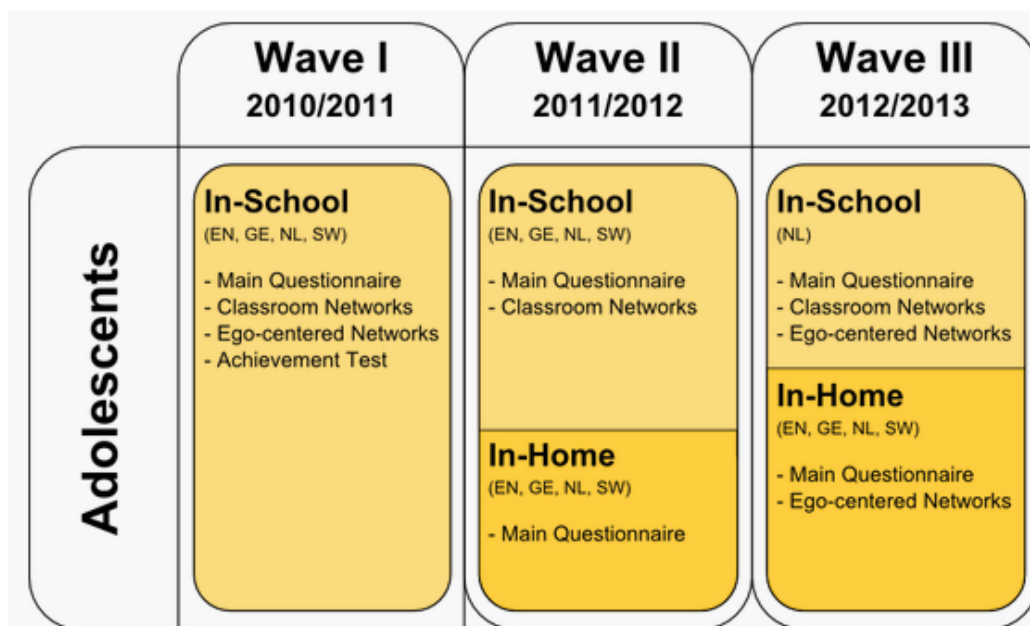
### 3.4.1 Identifier Variable

To uniquely identify the cases in the harmonised dataset, an identifier variable was included and named ‘H\_ID’. This variable applies to CILS4EU and NEPS SC4 cases and simply retains the values of the original identifier variable in each dataset – for the CILS4EU ‘youthid’ and for the NEPS SC4 ‘ID\_t’. As the identifier variables in the CILS4EU and NEPS SC4 differ in their number of digits (8 digits for ‘youthid’, 7 digits for ‘ID\_t’), the harmonised identifier ‘H\_ID’ allows for distinguishing between cases from CILS4EU and those from NEPS SC4.

### 3.4.2 Wave Indicators

In the first three waves of CILS4EU respondents were interviewed on a yearly basis. This wave structure is indicated by the variable ‘waveC’, which takes the value 1 for the year 2010/2011, 2 for 2011/2012, and 3 for 2012/2013 (see Figure 9 for an overview). In NEPS SC4, the data collection followed a different temporal pattern, with respondents being interviewed at different

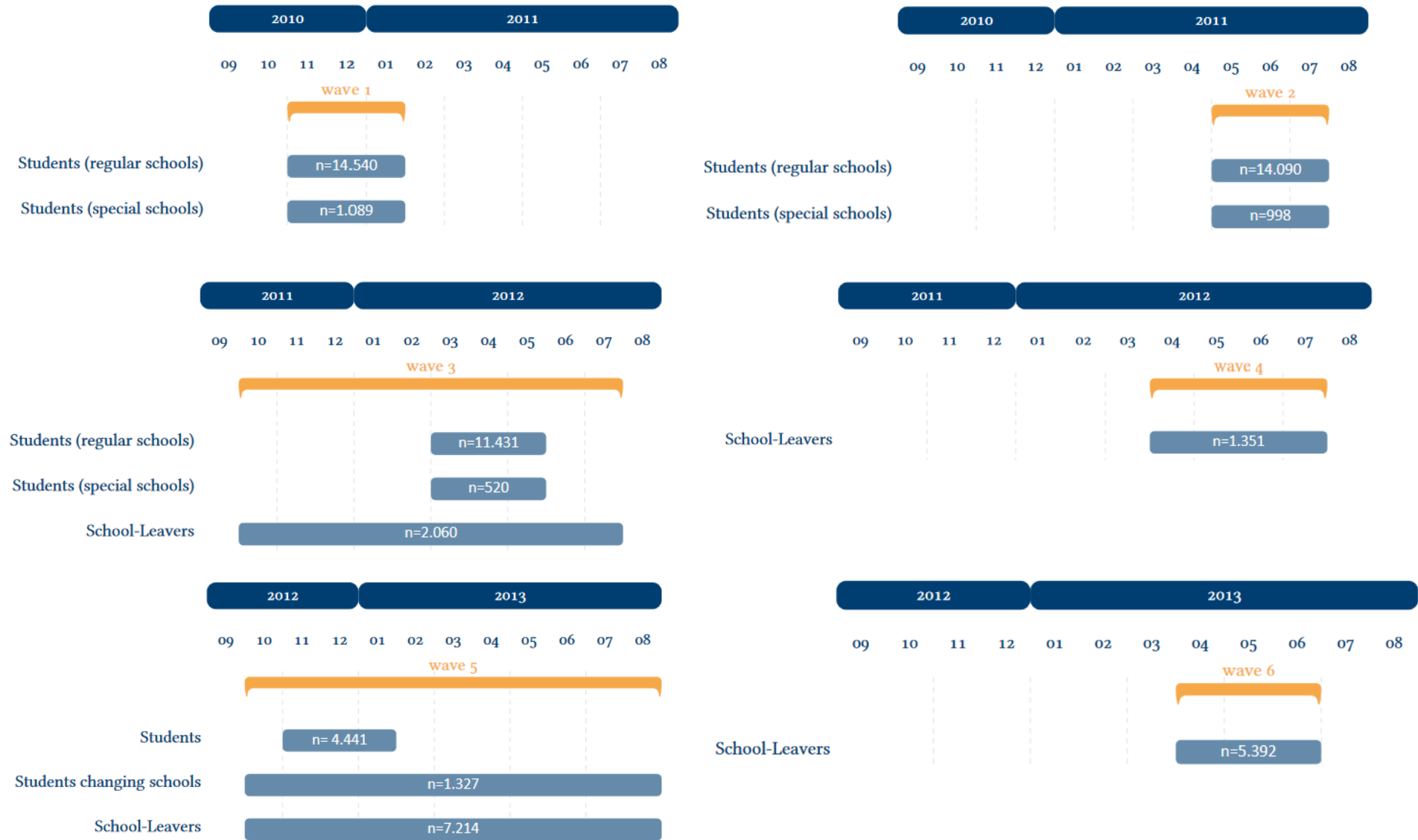
times depending on their status of students or school-leavers (please also refer to the NEPS SC4 documentation for detailed information; see also Figure 10 for an overview). The time frame of data collection in NEPS SC4 that matches the time frame of the first three waves of CILS4EU includes six instead of three waves. Wave 1, in which students were interviewed, was conducted from November 2010 to January 2011. Wave 2 followed in 2011 from May to July, again interviewing students. In wave 3, students were interviewed from March to May 2012 and school-leavers from October 2011 to July 2012. In this wave, it was possible that respondents were interviewed twice if their status changed: first as students and then as school-leavers. In wave 4, only school-leavers were interviewed in the period from April to July 2012. In wave 5, again students (November 2012 to January 2013) and school-leavers (October 2012 to August 2013) were interviewed. Also in this wave, it was possible that respondents were interviewed twice. Lastly, in wave 6 only school-leavers were interviewed (April to June 2013).



**Figure 9.** Wave Structure in CILS4EU: Wave 1–3

(See <https://www.cils4.eu>)





**Figure 10.** Wave Structure in NEPS SC4: Wave 1–6 (LifBi, 2021)

The difference in frequency of data collection between CILS4EU and NEPS SC4 as well as the distinction between students and school-leavers in NEPS SC4 (with different interview times and waves) prevents the construction of one uniform harmonised wave variable that exactly matches the CILS4EU and NEPS SC4 waves. Overall, waves 1, 3, and 5 in the NEPS SC4 match the three waves from CILS4EU best in terms of the time of data collection. Therefore, we recommend users conducting panel analyses with the harmonised dataset to compare these waves. To this end, we constructed a harmonised wave indicator ‘H\_wave2’, in which these respective CILS4EU and NEPS SC4 waves are merged.

One disadvantage of ‘H\_wave2’ is that it does not include waves 2, 4, and 6 from the NEPS SC4 – of which wave 4 and wave 6 are special school-leaver waves. This means that users with a particular interest in school-leavers miss their information to a certain extent if they rely on the ‘H\_wave2’ variable. Hence, although ‘H\_wave2’ provides the most reliable match of CILS4EU and NEPS SC4 waves, we provide an additional wave indicator for the harmonised dataset ‘H\_wave’. This wave indicator retains the original wave structure of the NEPS SC4 including six waves. However, as only three waves are available in CILS4EU for the respective time frame of data collection, waves 2, 4, and 6 from the ‘H\_wave’ variable include NEPS SC4 waves only. While overall only three waves can be used for panel analyses with the harmonised data, the ‘H\_wave’ variable allows users to decide more freely which NEPS SC4 waves to match with CILS waves. It thus becomes possible to compare the two school-leaver waves in NEPS SC4 (wave 4 and 6) with wave 2 and wave 3 in CILS4EU, respectively. Figure 11 provides an overview of ‘H\_wave2’ and ‘H\_wave’.

	2010/11		2011/12		2012/13	
<b>NEPS SC4</b>	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
<b>CILS4EU</b>	Wave 1		Wave 2		Wave 3	
<b>CILS4NEPS</b>						
<i>H_wave</i>	Wave 1	Wave 2 (NEPS only)	Wave 3	Wave 4 (NEPS only)	Wave 5	Wave 6 (NEPS only)
<i>N</i>	34,293	15,133	29,796	1,351	24,121	5,392
<i>H_wave2</i>	Wave 1	-	Wave 2	-	Wave 3	-
<i>N</i>	34,293		29,796		24,121	

**Figure 11.** Tabulation of the Two Harmonised Wave Indicators ‘H\_wave2’ and ‘H\_wave’

### 3.4.3 Dealing with Duplicate Cases in the Harmonised Dataset

As stated above, due to the wave structure in NEPS SC4, in which both students and school-leavers were interviewed in wave 3 and 5, it is possible that the same individuals were included twice in these waves – if they were first interviewed as students and changed their status to school-leaver during the data collection in the respective wave. In NEPS SC4, this does not pose problems because separate variables are included for students and school-leavers in the data (and different datasets). For instance, in wave 3 students' sex is recorded with the variable 't700031' whereas school-leavers' sex is recorded with the variable 't700001'. Yet, these duplicate interviews per wave pose problems for data harmonisation if both student and school-leaver variables are included in the same wave. In the example of respondents' sex, both variables the NEPS SC4 mentioned above form the harmonised item – which is necessary to not miss information from students or school-leavers. Now, if a respondent is interviewed both as a student and as a school-leaver in wave 3 answers are available for this person for both variables. If these answers are divergent, the question occurs which values are included in the harmonised variable. We dealt with this problem in the following way:

First, we checked for all cases in waves 3 and 5 in which variables from both students and school-leavers were included in the harmonised item whether the answers diverged. Please refer to Appendix B (Section 7.2) for an overview. Overlaps in answers occurred for a total of nine variables – which exclusively can be grouped into the harmonisation categories 'general information' ('sex'; 'day of birth, month'; 'day of birth, year') and 'household situation'. Second, for the variable 'general information', we checked whether a divergent answer might be erroneous. To this end, we compared the divergent answers with answers to the same variable in the waves prior to and after waves 3 and 5. If, for instance, a respondent stated 'January' as month of birth in waves 1 and 2, student wave 3 and wave 6 but 'July' in the school-leaver wave 3, we assumed that this answer was erroneous and used the value 'January' for the harmonisation of the variable 'day of birth, year'. However, for variables belonging to 'household situation', short-term changes in the household composition or living situation between the student and school-leaver interviews in wave 3 (and wave 5) might have taken place. Hence, we were not able to conduct a plausibility check for divergent answers here. Due to this, for all variables in the category 'household situation' with divergent answers, we always used the most recent answer (which is the school-leaver variable) for the construction of the harmonised variable – if this answer was valid (i.e. non-missing) However, if the value of this school-leaver variable was missing but the school-leaver variable contained a valid answer, we used the value of the student variable for the construction of the harmonised item.

### 3.5 Weighting

When combining the CILS4EU and NEPS SC4 datasets, it became necessary to adjust the existing weights. In principle, the two samples are non-disjoint, but schools were selected for the CILS4EU sample in such a way that there was no overlap with schools sampled in NEPS SC4. Nevertheless, the respective design weights were adjusted after combining the two samples by the statistical office at LIfBi. Please refer to Würbach and Aßmann (2023) for the technical report of the harmonised weights construction. In the harmonised dataset six harmonised weight variables are available:

w_t_CILS4NEP	Nonresponse adjusted joint panel entry weight for targets with panel consent (unstandardized)
w_t_CILS4NEPS_cal	Calibrated nonresponse adjusted joint panel entry weight for targets with panel consent (unstandardized)
w_t1_CILS4NEPS	Cross-sectional weight for targets participating in wave 1 (unstandardized)
w_t_CILS4NEPS_std	Nonresponse adjusted joint panel entry weight for targets with panel consent (standardized)
w_t_CILS4NEPS_cal_std	Calibrated nonresponse adjusted joint panel entry weight for targets with panel consent (standardized)
w_t1_CILS4NEPS_std	Cross-sectional weight for targets participating in wave 1 (standardized)

Which weights specifically are used for the analyses lies in the data users' decision, however it is advisable to use standardised weights in general. The harmonised weights include only German cases and are hence not available for respondents from the other three countries of the CLS4EU. Technically, the harmonised weights would allow for pooled analyses across the four countries, with the harmonised w\_t\_CILS4NEPS\_std being the most comparable to the CILS4EU house weight ('houwgt'). However, we advise users to estimate models separately per country and to compare coefficients. To do so, we constructed two additional weights for the harmonised dataset which include CILS4EU non-German cases only: 'w\_t\_CILS4EU\_std' and 'w\_t\_CILS4EU'. These weights are direct replicas of the CILS4EU 'houwgt' and 'totwgt' (please refer to the CILS4EU documentation material for detailed information on these weights).

w_t_CILS4EU_std	CILS4EU House weight (excluding German cases)
w_t_CILS4EU	CILS4EU Final Student weight (excluding German cases)

### **3.6 Information on Data Use**

Although we developed a stringent data harmonisation procedure and thoroughly checked the data before publication, errors in the harmonised dataset are not foreclosed. This is especially the case as ex-post harmonisation always comes with uncertainty in the resulting data. As stated above, even small deviations in the similarity between origin concepts can manifest themselves in concept mismatch and thus introduce bias in the dataset (Singh, 2020/2021). The extent to which dissimilarity between concepts is within limits or leads to such bias is difficult to gauge and no absolute certainty exists. The literature on data harmonisation suggests including a few checks in the analyses that make use of harmonised data (see, for example, Kolen and Brennan, 2014; Singh, 2021). For example, checks include testing whether the harmonised target items used in the analysis correlate with variables from the origin datasets in a way that would be expected from the theoretical literature (Singh, 2020; Kolen and Brennan, 2014). We do not go into detail here concerning these checks but strongly suggest that users inform themselves on the different ways to assess the robustness of their analyses when using the harmonised CILS4NEPS dataset.

## **4 Special Harmonised Variables**

### **4.1 Generational Status and Ethnic Origin**

The harmonisation of information on generational status and ethnic origin was particularly difficult because CILS4EU and NEPS follow different heuristics in defining these constructs (Dollmann, Jacob and Kalter, 2014; Olczyk, Will and Kristen, 2014).

The aim of the harmonisation process was to harmonise the variables ‘generational status’ and ‘ethnic origin’ in the two datasets using both the NEPS classification approach and the CILS4EU classification approach. Thus, a set of four variables was generated, with two variables each for generational status and ethnic origin – one for each classification approach. Although the underlying standard classification approaches in both datasets follow a similar logic, it is not feasible to use a 1:1 adaptation of the NEPS classification coding for the CILS4EU dataset and vice versa. This is due to dataset-specific differences in the value coding of the variable ethnic origin and varying definitions of the missing codes depending on the dataset.

#### 4.1.1 General remarks on the collection of information in CILS4EU and NEPS SC4

For the construction of ‘generational status’ and ‘ethnic origin’, both CILS4EU and NEPS rely on the countries of birth of respondents and their ancestors (i.e. their parents and grandparents). While respondents in NEPS reported this information for all their ancestors themselves, respondents in CILS4EU did not report the specific countries of birth for their grandparents. They did, however, indicate whether grandparents were born in the survey country or not, while information on grandparents was administered within parent interviews. Therefore, to define the countries of birth used for the construction of the variables ‘generational status’ and ‘ethnic origin’, CILS4EU first used any information available on the countries of birth of respondents’ parents and grandparents from the parent interviews. If a parent interview was not available or information was missing, information from adolescent interviews was used. Information on respondents’ own countries of birth was collected only in student interviews (see Dollmann, Jacob and Kalter, 2014 for more details).

Moreover, in CILS4EU, information on respondents’ and their ancestors’ countries of birth as well as age at migration was also collected in the second wave for all respondents (and partly also in the third wave). Thus, missing information on the various relevant variables in wave 1 could be substituted with that from wave 2 (and in wave 2 with wave 3 information, accordingly), and the ‘generational status’ and ‘ethnic origin’ were constructed again for wave 2 (and wave 3, accordingly). Information needed to construct the background variables was updated over time, and these variables differ between waves for some respondents (see Dollmann, Jacob and Kalter, 2016 for more details). In NEPS, in contrast, information about countries of birth was collected only in the first wave in which a respondent participated, and the ‘generational status’ and ‘ethnic origin’ were constructed only once per respondent.

When constructing the ‘generational status’ and ‘ethnic origin’ for the CILS4EU sample according to the NEPS classification approach, countries of birth were defined as described above, and both background variables were constructed with the most up-to-date information. In contrast, for the construction of the two variables for the NEPS sample according to the CILS4EU classification approach, the information provided by respondents in their first interview was used together with information from the wave 1 parent interview.<sup>3</sup>

---

<sup>3</sup> In N=49 cases, respondents received questions on their own country of birth as well as on their parents’ and grandparents’ countries of birth in two different waves. In accordance with the CILS4EU strategy, we prioritized the first observed information and filled missing and unclear values in with the second observed information.

Lastly, in CILS4EU, additional information was used to determine immigrant background, namely information on children's and parents' ethnic identity, nationality, and children's self-reported information on immigrant background status. For constructing 'generational status' and 'ethnic origin' for NEPS SC4 respondents according to the CILS4EU classification approach only children's and parents' nationality was used since the other information was not available.

**Important note:** In the final combined data set, all data lines of CILS4EU respondents (i.e. waves) contain the same information on the harmonised variables for 'generation status' and 'ethnic origin' as defined according to the NEPS SC4 classification approach. For NEPS SC4 respondents, these variables contain the original information (and therefore, only contain information in the first interview; see description above). All data lines of NEPS SC4 respondents (i.e. waves) contain the same information on the harmonised variables for 'generation status' and 'ethnic origin' as defined according to the CILS4EU classification approach for all waves. For CILS4EU respondents, these variables contain the original information (and therefore, contain potentially updated and thus time-varying information; see description above).

#### 4.1.2 Coding Strategies: Generational Status

In CILS4EU, generational status is classified using a systematic top-down approach: First, the child's country of birth is considered, then the parents' country of birth and finally the grandparents' country of birth (see also Dollmann, Jacob and Kalter, 2014). Thus, information on the country of birth of seven actors is used (child, two parents and four grandparents). This approach allows for a fine-grained distinction of generational status, distinguishing between the 1<sup>st</sup>, 1.25<sup>th</sup>, 1.5<sup>th</sup>, 1.75<sup>th</sup>, 2<sup>nd</sup>, 2.5<sup>th</sup>, 2.75<sup>th</sup>, interethnic 2<sup>nd</sup>, 3<sup>rd</sup>, 3.25<sup>th</sup>, 3.5<sup>th</sup>, interethnic 3<sup>rd</sup>, and 3.75<sup>th</sup> generation and natives.

The NEPS SC4 generational status variable is based on the information on the country of birth of the target person and their parents and grandparents, which is in line with the CILS4EU standard classification approach. NEPS SC4 allows for a fine-grained differentiation of generational status as well, distinguishing between the 1<sup>st</sup>, 1.5<sup>th</sup>, 2<sup>nd</sup>, 2.25<sup>th</sup>, 2.5<sup>th</sup>, 2.75<sup>th</sup>, 3<sup>rd</sup>, 3.25<sup>th</sup>, 3.5<sup>th</sup> generation, and the majority. Table 4 Column B) provides a definition of the respective NEPS SC4 generational status classification (see also Olczyk, Will and Kristen, 2014). To facilitate the comparability of the two approaches (see also Table 4 Column C)), the NEPS SC4 approach was graphically adapted to the country-of-birth ancestry scheme proposed in the CILS4EU classification approach (see Dollmann, Jacob and Kalter, 2014: 10).

Table 4 provides a comparison of the classification approaches for generational status in CILS4EU and NEPS SC4. Columns A) and B) provide a dataset-specific definition (Dollmann, Jacob and Kalter, 2014: 10; Olczyk, Will and Kristen, 2014: 8). A graph of the country-of-birth ancestry scheme complements the definition of the respective generational status. Light grey rectangles represent actors born in the survey country; dark grey ones represent actors born outside the survey country. White rectangles are used to indicate that the country of birth of the actor is irrelevant for assessing the ‘ancestral distance from the point of arrival’ (Alba, 1988: 213) and therefore for defining the generational status of the child. Column C) depicts commonalities, similarities, and equivalent classifications as well as differences between these two dataset-specific approaches.

### Overlap

In general, both CILS4EU and NEPS SC4 are based on the systematic classification of generational status that follows the top-down approach. In CILS4EU, however, the top-down approach is used more strictly up to the 3<sup>rd</sup> generation than in NEPS SC4.

There are four identical assignments of generational status in both approaches: 1<sup>st</sup> generation, 2<sup>nd</sup> generation, 3<sup>rd</sup> generation and 3.25<sup>th</sup> generation (see also Table 4 Column C).

Accordingly, CILS4EU and NEPS SC4 define first-generation migrants as foreign-born persons who themselves migrated to the survey country. In addition, the age at migration is considered in both approaches. However, the approaches use different age categories, resulting in discrepancies within the subclassification of the 1<sup>st</sup> generation (for more details, see ‘Differences’).

Both CILS4EU and NEPS SC4 classify second-generation migrants as persons who were born in the survey country with both parents born abroad, regardless of the country of birth of the grandparents. Additionally, in both approaches, the 3<sup>rd</sup> generation comprises target persons who were born in the survey country with both parents also born in survey country, but in this group, all grandparents are born abroad. In this context, the CILS4EU and NEPS SC4 approaches also overlap in terms of subclassification of the 3.25<sup>th</sup> generation. Thus, in both approaches target persons are assigned to the 3.25<sup>th</sup> generation if they were born in the survey country with both parents born in the survey country as well, while having three foreign-born grandparents.

### Differences

There are two major differences between the CILS4EU and NEPS SC4 classification approaches. First, the CILS4EU classification differentiates the generational status up to the



3.75<sup>th</sup> generation. Children born in the survey country with both parents born in the survey country as well but with only one foreign-born grandparent represent the 3.75<sup>th</sup> generation. NEPS SC4 does not include this subclassification anymore, already classifying this group as majority. This is also empirically shown in Table 5. The equivalent category for the NEPS majority group is the native group in CILS4EU. In CILS4EU, natives are defined as target persons who were born in the survey country and whose parents and grandparents were born in the survey country.

Second, in CILS4EU, the subclassifications of the 1<sup>st</sup> generation are more strongly differentiated by several age groups at migration. In NEPS SC4, only the age at school entry, i.e., migration before or after the age of 6, is taken into account; CILS4EU further distinguishes between migration at ages 0–5, 6–10, and older than 11. Consequently, the CILS4EU approach includes two further subclassifications of the 1<sup>st</sup> generation, i.e., the 1.25<sup>th</sup> and 1.75<sup>th</sup> generation.

Another striking difference between the two approaches is the CILS4EU-specific subclassification of the 2<sup>nd</sup> and 3<sup>rd</sup> generation, labelled interethnic 2<sup>nd</sup> and interethnic 3<sup>rd</sup>, respectively. This distinction is not used in the NEPS approach.

The category interethnic 2<sup>nd</sup> generation denotes target persons who have one parent who is a first-generation migrant and one parent who was born in the survey country (interethnic partnership). An equivalent classification is the NEPS-specific 2.75<sup>th</sup> generation. This is also empirically shown in Table 5.

Analogously, the variable interethnic 3<sup>rd</sup> generation captures children with one second-generation parent and one parent whose parents were born in the survey country. An equivalent is the NEPS-specific 3.5<sup>th</sup> generation.

**Important note:** When cross-tabulating the different approaches, one has to keep in mind that the approaches make use of different sources of information (e.g., prioritizing parent information about ancestors over children information in the CILS4EU approach). As is apparent from Table 4, this leads to several inconsistencies between the variables. Such differences might occur, for instance, when the child reports having foreign-born parents, while the interviewed parent reports being German born to foreign-born parents.

**Table 4.** Overview of Dataset-Specific\* Classification Approaches for Generational Status

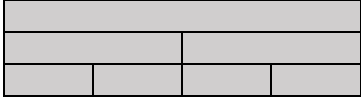
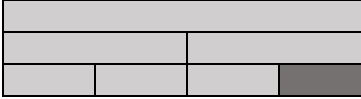
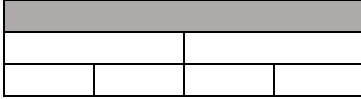
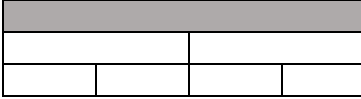
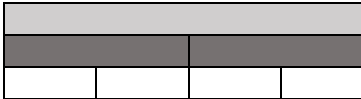
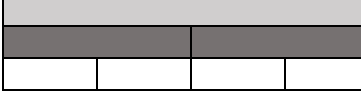
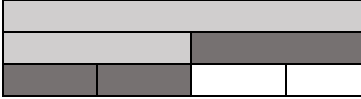
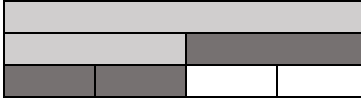
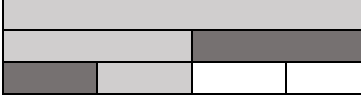
Gen. Status	A) CILS4EU	B) NEPS	C) Differences/overlaps between NEPS and CILS4EU
Natives/majority	All actors born in survey country (SC) 	Target person and parents born in Germany; at most one grandparent (if any) born abroad 	In some cases, the NEPS approach classifies individuals as majority, still covered by the 3.75 <sup>th</sup> generation according to the CILS4EU approach.
1 <sup>st</sup>	Child born abroad, irrespective of the countries of birth of the ancestors 	Target person born abroad and immigrated after the age of 6 	Identical
1.25 <sup>th</sup>	Migration at age 11+		NEPS equivalent: 1 <sup>st</sup> generation
1.5 <sup>th</sup>	Migration at age 6–10	Target person born abroad and immigrated before the age of 6	
1.75 <sup>th</sup>	Migration at age 0–5		NEPS equivalent: 1.5 <sup>th</sup> generation
2 <sup>nd</sup>	Child born in SC and both parents born abroad, irrespective of the countries of birth of the grandparents 	Target person born in Germany and both parents born abroad 	Identical
2.25 <sup>th</sup>		Target person born in Germany with one parent born abroad and the other in Germany; parents of the latter both born abroad 	CILS4EU equivalent: 2 <sup>nd</sup> generation
2.5 <sup>th</sup>	Child born in SC with one parent also born in SC and the other born abroad; parents of the former both born abroad 	Target person born in Germany with one parent born abroad and the other in Germany; one parent of the latter born abroad 	CILS4EU equivalent: 2.75 <sup>th</sup> generation

Table 4. Continued

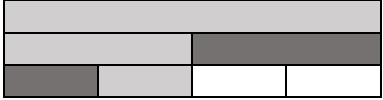
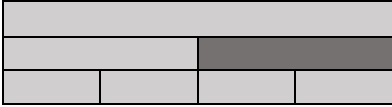
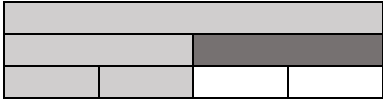
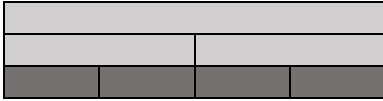
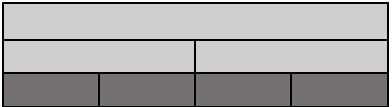

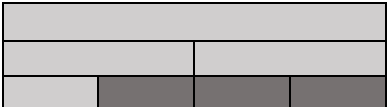
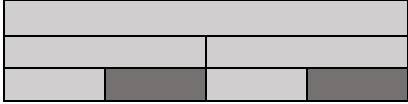
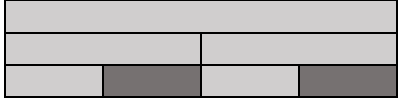
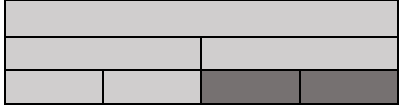
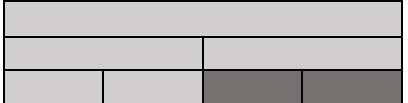
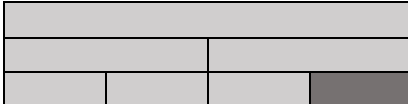
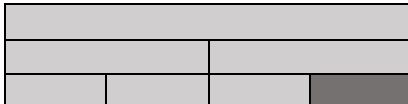
2.75 <sup>th</sup>	<p>Child born in SC with one parent also born in SC and the other abroad; one parent of the former also born in SC and the other born abroad, irrespective of the countries of birth of the foreign-born parent's parents</p> 	<p>Target person born in Germany with one parent born abroad and the other in Germany; no grandparents born abroad</p> 	<p>Diverging classification between NEPS &amp; CILS4EU, CILS4EU equivalent: interethnic 2<sup>nd</sup></p>
<i>Interethnic 2<sup>nd</sup></i>	<p>Child born in SC with one parent also born in SC and the other abroad; parents of the former both born in SC, irrespective of the countries of birth of the foreign-born parent's parents (being a first-generation immigrant)</p> 		
3 <sup>rd</sup>	<p>Child and parents born in SC; all grandparents born abroad</p> 	<p>Target person and parents born in Germany; all grandparents born abroad</p> 	Identical
3.25 <sup>th</sup>	<p>Child and parents born in SC; three grandparents born abroad and one in SC</p> 	<p>Target person and parents born in Germany; three grandparents born abroad</p> 	Identical

Table 4. Continued

3.5 <sup>th</sup>	<p>Child born in SC with both parents also born in SC, both of whom have one parent born abroad and one parent born in SC</p> 	<p>Target person and parents born in Germany; two grandparents born abroad</p>  	
<i>Interethnic 3<sup>rd</sup></i>	<p>Child born in SC with both parents also born in SC; two grandparents born in SC and the other two abroad</p> <p>This category therefore comprises children descending from a relationship between a 2<sup>nd</sup>-generation parent and a parent whose parents also were both born in SC.</p> 		In NEPS partly covered by the NEPS-specific 3.5 <sup>th</sup> generation.
3.75 <sup>th</sup>	<p>Child born in SC with both parents also born in SC and one grandparent born abroad</p> <p>In this sense, the 3.75<sup>th</sup> generation is to some degree comparable to the interethnic 3<sup>rd</sup> generation, as the child has one parent whose parents were both born in SC and one parent who is from the 2.5<sup>th</sup>, 2.75<sup>th</sup>, or interethnic 2<sup>nd</sup> generation.</p> 	<p>Target person and parents born in Germany; one grandparent born abroad</p> 	Identical

Note: \*Definitions of generational status within the dataset-specific classification in columns A) and B) are based on the original definition used in the respective survey (cf. Dollmann, Jacob and Kalter, 2014; Olczyk, Will and Kristen, 2014).

**Table 5.** Cross-Tabulation of the Variable Generational Status for CILS4EU and NEPS SC in Wave 1 According to the CILS4EU Approach (Rows) and According to the NEPS Approach (Columns) (row percentages)

CILS4EU approach (gen. status)	NEPS approach (gen. status)												Total
	Not determinable	Majority	1 <sup>st</sup>	1.5 <sup>th</sup>	2 <sup>nd</sup>	2.25 <sup>th</sup>	2.5 <sup>th</sup>	2.75 <sup>th</sup>	3 <sup>rd</sup>	3.25 <sup>th</sup>	3.5 <sup>th</sup>	3.75 <sup>th</sup>	
1.25 <sup>th</sup>	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
1.5 <sup>th</sup>	0.00	0.00	78.35	21.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
1.75 <sup>th</sup>	0.00	0.00	0.00	99.62	0.23	0.00	0.00	0.15	0.00	0.00	0.00	0.00	100.00
1 <sup>st</sup> gen. missing migration age	0.00	0.84	83.97	12.66	0.84	0.00	0.00	0.84	0.00	0.00	0.42	0.42	100.00
2 <sup>nd</sup>	0.00	0.23	0.00	0.00	95.40	1.78	0.46	1.21	0.56	0.00	0.31	0.06	100.00
2.5 <sup>th</sup>	0.00	0.23	0.00	0.00	7.83	69.86	4.56	15.07	1.99	0.00	0.47	0.00	100.00
2.75 <sup>th</sup>	0.00	0.63	0.00	0.00	4.69	1.88	83.75	6.56	0.00	0.63	0.94	0.94	100.00
Interethnic 2 <sup>nd</sup>	0.00	1.85	0.00	0.00	0.65	0.42	2.09	93.14	0.00	0.00	1.20	0.65	100.00
3 <sup>rd</sup>	0.00	3.82	0.00	0.00	1.15	5.34	0.00	0.00	59.54	3.05	17.56	9.54	100.00
3.25 <sup>th</sup>	0.00	1.69	0.00	0.00	0.00	1.69	3.39	0.00	0.00	72.03	19.49	1.69	100.00
3.5 <sup>th</sup>	0.00	13.87	0.00	0.00	0.00	0.00	1.09	1.09	0.00	0.73	47.81	35.40	100.00
Interethnic 3 <sup>rd</sup>	0.00	5.10	0.00	0.00	0.00	0.24	0.24	4.63	0.71	0.95	72.12	16.01	100.00
3.75 <sup>th</sup>	0.00	8.74	0.00	0.00	0.00	0.00	0.44	0.92	0.00	0.22	1.79	87.89	100.00
Native	0.01	98.10	0.00	0.00	0.04	0.00	0.00	0.30	0.01	0.01	0.22	1.31	100.00
<i>Missing information, but immigrant background</i>													
Parents foreign-born, no info on child	2.17	0.00	0.00	4.35	86.96	2.17	2.17	2.17	0.00	0.00	0.00	0.00	100.00
Child native-born, no info on parents, grandparents foreign-born	0.00	3.45	0.00	0.00	20.69	6.90	0.00	0.00	37.93	3.45	13.79	13.79	100.00
Child native-born, at least one ancestor foreign-born	0.00	98.11	0.00	0.00	1.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
<i>Missing information, immigrant background unclear</i>													
Child native-born, no info on parents and grandparents	0.00	91.87	0.00	0.00	4.07	0.00	0.00	4.07	0.00	0.00	0.00	0.00	100.00
Child and parents native-born, no info on grandparents	0.00	98.68	0.00	0.00	0.00	0.00	0.00	1.32	0.00	0.00	0.00	0.00	100.00
No information on any actor	88.89	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

*Note.* Row percentages are displayed.

### 4.1.3 Coding Strategies: Ethnic Origin

Previous research provides numerous approaches to determine ethnicity, using various indicators. Commonly, self-subscribed ethnic identity, nationality, language use and countries of birth of the ancestors are the most relevant and most often used indicators (e.g. Gresch and Kristen, 2011; Jacobs *et al.*, 2009).

In CILS4EU, ethnicity is defined by parentage, as this can also be directly linked to generational status. CILS4EU includes the specific countries of birth of the children, their parents and their grandparents. However, the ethnic origin assignment is started at the grandparent level to define the family's origin before migration to the respective survey country. Accordingly, a bottom-up approach is used for the specification of ethnic origin – as opposed to the top-down approach for the generational status variable. This approach is applied to all persons classified as migrants (cf. generational status) as well as to all persons with missing information for which a migration background cannot be excluded (for a detailed overview of the classification approach regarding ethnic origin see Dollmann, Jacob and Kalter (2014: 24–26).

In NEPS, the starting point of the assignment to a certain ethnic group is based on the information about the country of birth and the generational status. As opposed to CILS4EU, in NEPS information on ethnic origin is coded specifically only for countries of origin that 'have significantly shaped and are still shaping the (contemporary) history of migration to Germany' (Olczyk, Will and Kristen, 2014: 13). Thus, only the five largest migrant groups living in Germany are explicitly listed (by country of origin), namely Germany, Italy, Poland, Romania and Turkey. All other countries of origin are categorised according to geographical criteria,<sup>4</sup> for example in Northern and Western Europe, North America or Former Soviet Union (see also Olczyk, Will and Kristen, 2014: 13, Table 3). Furthermore, NEPS SC4 classifies ethnic origin at the level of the target persons and their generational status. The specific rules of assignment to a certain ethnic origin are specified in Olczyk, Will and Kristen (2014: 13–15).

## 4.2 Tracking Variables

To identify respondents' participation status, we constructed two tracking variables – one for CILS4EU ('trackingC') and one for NEPS SC4 ('trackingN'), which also allow for the

---

<sup>4</sup> I.e. Former Soviet Union, Central and South America, Caribbean, Northern and Western Europe, North America, Oceania/Polynesia, Other Middle East and North Africa, Other Africa, Other Asia, Other Central and Eastern Europe, and Other Southern Europe.

construction of a balanced panel for the harmonised dataset (which we describe below). The do-file ‘4\_Tracking Variables’ in the documentation material includes the code for the construction of the tracking variables.

For the CILS4EU variable, several temporary variables had to be created first: First, the wave number was squared; second, these values were summed up within each respondent ID. This resulted in values that indicate in which CILS4EU waves the respondent participated. The values 1, 2, and 3 indicate that the respondent only participated in the respective wave. The values 12, for instance, indicates participation in wave 1 and 2. The value 123 represents participation in all three CILS4EU waves.

For the NEPS SC4 tracking variable, temporary variables also had to be created first: one variable per wave including the respective wave number, which was then applied to all observations of the same respondent ID. Subsequently, these individual wave numbers were strung together. This resulted in the same format as the CILS4EU tracking variable has. Hence, the value 23 corresponds to participation in wave 2 and wave 3, while the value ‘123456’ indicates participation in all six waves (which is unlikely due to the distinction between student and school-leaver waves).

## **5 Construction of Filter Variables**

### **5.1 Balanced Panel**

Based on the CILS4EU and NEPS SC4 tracking variables explained above, we included a filter variable in the harmonised dataset that allowed us to construct the balanced panel ‘filter\_balanced’. Based on the harmonised wave variable ‘H\_wave2’, this filter takes the value 1 for all respondents who participated in all three harmonised waves (i.e. CILS4EU wave 1, 2, and 3; NEPS SC4 wave 1, 3, and 5). Please refer to the do-file ‘6\_Balanced Panel Flag’ in the documentation for the code to construct this flag variable.

### **5.2 Selection of Items Classified as Unproblematic versus More Complicated**

As described in Section 3, we classified the harmonised variables as unproblematic and more complicated – a classification that correlates with the harmonisation strategy. All variables classified as unproblematic are manifest constructs and allowed for a simple matching of answer categories. Hence, concept mismatch and resulting biases in the analyses are unlikely for these variables. More complicated variables mostly comprise latent constructs, for which linear equating was necessary. Although we carefully assessed concept similarity for these

items before equating, biases caused by their harmonisation cannot be ruled out (see also Section 3.6). Because of this, we provide users with a simple way to exclude all more complicated variables from the dataset. To do so, a characteristic called `complicated` was created for all variables. This characteristic takes the value 1 for more complicated variables and 0 for all unproblematic variables. Based on this characteristic, a loop was created in the do-file ‘10\_Flags\_Equating\_Wellenabstand’, which allows for removing all more complicated variables from the dataset. This do-file is provided in the documentation of the harmonised dataset and can be run by users.

### 5.3 Linear Equating: Differences in Year-Overlap

As described in Section 3.3.3.2, one prerequisite for the linear equating procedure to produce reliable results and to avoid temporal and spatial (e.g. cultural) differences is that the samples from each source dataset refer to the same population and that items that are equated were surveyed within the same time frame (Kolen and Brennan, 2014; Singh, 2020). We fulfil this prerequisite except for six variables (see Appendix C Section 7.3 for a list) – for which the difference in the survey year they were collected between CILS4EU and NEPS SC4 is one year or more. To enable users to exclude these variables from the analyses, we constructed a code similar to the one for unproblematic and more complicated items explained above.

Again, a characteristic, called `equate`, was created. This characteristic takes the value 1 if the CILS4EU and the NEPS SC4 variables that form the harmonised equated item were collected within a time difference of one year, and the value 4 if there is a difference of four years between the CILS4EU and NEPS SC4 data collection. The characteristic only takes the values 1 or 4, as no other year distances occurred for the equated items in the data collection. Based on this characteristic, a loop was created in the do-file ‘10\_Flags\_Equating\_Wellenabstand’, which allows for removing all harmonised equated items for which the source variables of CILS4EU and NEPS SC4 were not collected in the same years.

## 6 Missing Values

In the construction of the harmonised dataset, a harmonised missing scheme combining the missing values of CILS4EU and NEPS SC4 was created. Figure 12 and Figure 13 represent the individual missing schemes in CILS4EU and NEPS SC4 respectively. These figures show that while there is overlap in missing codes between CILS4EU and NEPS SC4, overall missing codes differ between the two datasets. Furthermore, while filters in the PAPI mode allow for



self-assignment of respondents in NEPS SC4, leading to potential logical inconsistencies in the answers, this is not the case in CILS4EU.

Code	Label	Description
-99	Don't know	Only if this option was provided in the questionnaire
-88	No answer	No box ticked/no answer given
-77	Not applicable	Not answered due to filter question
-66	Question not asked	Not asked in a country/mode
-55	Other missing	Answers that could not be coded, fun answers
-44	Interrupted Interview	Only for telephone and online modes
-33	Not available in reduced version	Answers not available in the download version

**Figure 12.** Missing Codes in CILS4EU (CILS4EU, 2016)

Code	Missing
<b>Item nonresponse</b>	
-97	refused
-98	don't know
-95	implausible value
-94	not reached (only applicable for competence tests)
-5/-6/-20,...,-29	item-specific missing with informative value labels
<b>Not applicable</b>	
-54	missing by design (mostly: not included in sample-specific instrument of this wave)
-93	does not apply
.	filtered / system missing (in CATI/CAPI mode)
-90	unknown missing
-99	filtered (in PAPI mode)
<b>Edition missings (recoded into missing)</b>	
-52	implausible value removed
-53	Anonymized
-55	not determinable
-56	Not participated

**Figure 13.** Missing Codes in NEPS SC4 (Skopek, Pink and Bela, 2013)

Based on these differences, we decided to create a missing scheme for the harmonised dataset in which missings in CILS4EU and NEPS SC4 are not combined but listed separately. This ensures that all types of missings can be considered and, at the same time, that no errors occur when combining missing codes of the two datasets that have a similar label but follow a different underlying logic. One exception is the missing label 'don't know', which is identical

in CILS4EU and NEPS SC4 and therefore combined in the harmonised dataset. **Fehler! Ungültiger Eigenverweis auf Textmarke.** displays the harmonised missing scheme. It shows which missing codes belong to which source dataset. The missing values ‘not determinable’, ‘filtered’ and ‘don’t know’ in NEPS SC4 were recoded to prevent overlap with missing codes in CILS4EU that hold the same value but refer to a different missing type.

**Table 6.** Missing Codes in the Harmonised Dataset

Dataset	Value	Label	Recoding
NEPS SC4	-5/-6/-20,...,-29	Item-specific missing with informative value label	
CILS4EU	-33	Not available in RV	
CILS4EU	-44	Interrupted interview	
NEPS SC4	-52	Implausible value removed	
NEPS SC4	-53	Anonymised	
NEPS SC4	-54	Missing by design	
CILS4EU	-55	Other missing	
NEPS SC4	-56	Not participated	
NEPS SC4	-57	Not determinable	Previously -55 in NEPS SC4
CILS4EU	-66	Question not asked	
CILS4EU	-77	Not applicable	
CILS4EU	-88	No answer	
NEPS SC4	-90	Unknown missing	
NEPS SC4	-92	Question erroneously not asked	
NEPS SC4	-93	Does not apply	
NEPS SC4	-94	Not reached	
NEPS SC4	-95	Implausible value	
NEPS SC4	-97	Refused	
NEPS SC4	-98	Filtered	Previously -99 in NEPS SC4
Combined	-99	Don't know	Previously -98 in NEPS SC4
NEPS SC4	.	Filtered/system missing	

## 7 Appendix

### 7.1 A – Documentation of Changes in the CILS4EU Data for the Converting from a Wide to a Long Data Format Converting

H: s_csit2CS		CILS4EU W2: y2_s_csit2CS		CILS4EU W3: y3_s_csit2CS	
<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>
1	ge: lower secondary school (hauptschule)	1	GE: Lower secondary school (Hauptschule)	1	GE: Lower secondary school (Hauptschule)
2	ge: intermediate secondary school (realschule)	2	GE: Intermediate secondary school (Realschule)	2	GE: Intermediate secondary school (Realschule)
3	ge: upper secondary school (realschule plus)	-	-	3	GE: Upper secondary school (Realschule plus)
4	ge: upper secondary school (gymnasium)	<b>3</b>	GE: Upper secondary school (Gymnasium)	4	GE: Upper secondary school (Gymnasium)
5	ge: comprehensive school (integrierte gesamtschule)	<b>4</b>	GE: Comprehensive school (Integrierte Gesamtschule)	5	GE: Comprehensive school (Integrierte Gesamtschule)
6	ge: combined lower, intermediate and upper secondary school (kooperative gesamtschule)	<b>5</b>	GE: Combined lower, intermediate and upper secondary school (Kooperative Gesamtschule)	6	GE: Combined lower, intermediate and upper secondary school (Kooperative Gesamtschule)
7	ge: higher secondary vocational school (fachoberschule)	<b>6</b>	GE: Higher secondary vocational school (Fachoberschule)	7	GE: Higher secondary vocational school (Fachoberschule)
8	ge: combined lower and intermediate secondary school (mittelschule)	<b>7</b>	GE: Combined lower and intermediate secondary school (Mittelschule)	8	GE: Combined lower and intermediate secondary school (Mittelschule)
9	ge: combined lower and intermediate secondary school (regelschule)	<b>8</b>	GE: Combined lower and intermediate secondary school (Regelschule)	9	GE: Combined lower and intermediate secondary school (Regelschule)
10	ge: combined lower and intermediate secondary school (sekundarschule)	<b>9</b>	GE: Combined lower and intermediate secondary school (Sekundarschule)	10	GE: Combined lower and intermediate secondary school (Sekundarschule)
11	ge: combined lower and intermediate secondary school (haupt-realschule)	<b>10</b>	GE: Combined lower and intermediate secondary school (Haupt-Realschule)	11	GE: Combined lower and intermediate secondary school (Haupt-Realschule)
12	ge: school for special needs (foerderschule)	<b>11</b>	GE: School for special needs (Förderschule)	12	GE: School for special needs (Förderschule)

13	ge: rudolf-steiner school (waldorfschule)	<b>12</b>	GE: Rudolf-Steiner school (Waldorfschule)	13	GE: Rudolf-Steiner school (Waldorfschule)
14	ge: vocational school (berufsschule)	<b>13</b>	GE: Vocational school (Berufsschule)	14	GE: Vocational school (Berufsschule)
15	ge: full-time vocational school (berufsfachschule)	<b>14</b>	GE: Full-time vocational school (Berufsfachschule)	15	GE: Full-time vocational school (Berufsfachschule)
16	ge: higher full-time vocational school (hoehere berufsfachschule)	<b>15</b>	GE: Higher full-time vocational school (Höhere Berufsfachschule)	16	GE: Higher full-time vocational school (Höhere Berufsfachschule)
17	ge: commercial school (handelsschule)	<b>16</b>	GE: Commercial school (Handelsschule)	17	GE: Commercial school (Handelsschule)
18	ge: higher commercial school (hoehere handelsschule)	<b>17</b>	GE: Higher commercial school (Höhere Handelsschule)	18	GE: Higher commercial school (Höhere Handelsschule)
19	ge: other school type	<b>18</b>	GE: Other school type	<b>19</b>	GE: Other general educational school
20	nl: vmbo-basis	<b>19</b>	NL: vmbo-basis	<b>20</b>	GE: Other vocational school
21	nl: vmbo-kader	<b>20</b>	NL: vmbo-kader	<b>21</b>	NL: vmbo-basis
22	nl: vmbo-gt	<b>21</b>	NL: vmbo-gt	<b>22</b>	NL: vmbo-kader
23	nl: vmbo-t	<b>22</b>	NL: vmbo-t	<b>23</b>	NL: vmbo-gt
24	nl: havo	<b>23</b>	NL: havo	<b>24</b>	NL: vmbo-t
25	nl: vwo	<b>24</b>	NL: vwo	<b>25</b>	NL: havo
26	nl: gymnasium	<b>25</b>	NL: gymnasium	<b>26</b>	NL: vwo
27	nl: other school type	<b>26</b>	NL: Other school type	<b>27</b>	NL: gymnasium
28	sw: academic programme preparing for higher education	-	-	-	-
29	sw: vocational programme: school- located training	-	-	28	SW: Academic programme preparing for higher education
30	sw: vocational programme: workplace- based training	-	-	29	SW: Vocational programme: school- located training
31	sw: introductory programme: preparatory course	-	-	30	SW: Vocational programme: workplace- based training
32	sw: introductory programme: programme-oriented individual selection	-	-	31	SW: Introductory programme: preparatory course
				32	SW: Introductory programme: programmeoriented individual selection

33	sw: introductory programme: vocational introduction	-	-	33	SW: Introductory programme: vocational introduction
34	sw: introductory programme: individual alternative	-	-	34	SW: Introductory programme: individual alternative
35	sw: introductory programme: language introduction	-	-	35	SW: Introductory programme: language introduction
36	sw: other school type	-	-	36	SW: Other school type

H: rel1		CILS4EU W1: y1_rel1		CILS4EU W2: y2_rel1		CILS4EU W3: y3_rel1	
Value	Label	Value	Label	Value	Label	Value	Label
1	no religion	1	No religion	1	No religion	1	No religion
2	buddhism	2	Buddhism	2	Buddhism	2	Buddhism
3	christianity	3	Christianity	3	Christianity	3	Christianity
						<b>6</b>	Christianity: Other
4	christianity: catholic	4	Christianity: Catholic	4	Christianity: Catholic	4	Christianity: Catholic
5	christianity: protestant	5	Christianity: Protestant	5	Christianity: Protestant	5	Christianity: Protestant
6	hinduism	6	Hinduism	6	Hinduism	<b>7</b>	Hinduism
7	islam	7	Islam	7	Islam	<b>8</b>	Islam
8	judaism	8	Judaism	8	Judaism	<b>9</b>	Judaism
9	sikhism	9	Sikhism	9	Sikhism	<b>10</b>	Sikhism
10	other religion	10	Other religion	10	Other religion	<b>11</b>	Other religion

<b>H: s_gradeCS</b>		<b>CILS4EU W2: y2_s_gradeCS</b>		<b>CILS4EU W3: y3_s_gradeCS</b>	
<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>
1	ge: 9th grade	1	GE: 9th grade	1	GE: 9th grade
2	ge: 10th grade	2	GE: 10th grade	2	GE: 10th grade
3	ge: 11th grade	3	GE: 11th grade	3	GE: 11th grade
4	ge: no grade	4	GE: No grade	4	GE: No grade
5	ge: other grade	5	GE: Other grade	5	GE: Other grade
6	nl: 3rd grade	6	NL: 3rd grade	-	-
7	nl: 4th grade	7	NL: 4th grade	<b>6</b>	NL: 4th grade
8	nl: 5th grade	8	NL: 5th grade	<b>7</b>	NL: 5th grade
9	nl: other grade	9	NL: Other grade	<b>8</b>	NL: Other grade

<b>H: planCS</b>		<b>CILS4EU W2: y2_planCS</b>		<b>CILS4EU W3: y3_planCS</b>	
<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>	<i>Value</i>	<i>Label</i>
1	en: stay on the school you are at now	1	EN: Stay on in the school you are at now	1	EN: Stay on in the same school or college
2	en: move to a different school	2	EN: Move to a different school	2	EN: Move to a different school or college
3	en: move to a sixth form college	3	EN: Move to a sixth form college	-	-
4	en: move to a college of further education	4	EN: Move to a college of further education	-	-
5	en: leave school and get a full-time job	5	EN: Leave school and get a full-time job	<b>3</b>	EN: Work in a full-time job
6	en: leave school and start an apprenticeship	6	EN: Leave school and start an apprenticeship	<b>4</b>	EN: Complete an apprenticeship or workrelated training
7	en: internship	-	-	<b>5</b>	EN: Complete an internship
8	en: something else	<b>7</b>	EN: Something else	<b>6</b>	EN: Something else
9	ge: stay on in school and get degree from intermediate secondary school	<b>8</b>	GE: Stay on in school and get degree from intermediate secondary school	-	-
10	ge: stay on in school and get degree from upper secondary (vocational) school	<b>9</b>	GE: Stay on in school and get degree from upper secondary (vocational) school	-	-
11	ge: vocational preparation year	<b>10</b>	GE: Vocational preparation year	-	-
12	ge: full-time work	<b>11</b>	GE: Full-time work	-	-
13	ge: apprenticeship	<b>12</b>	GE: Apprenticeship	-	-
14	ge: internship	<b>13</b>	GE: Internship	-	-
15	ge: something else	<b>14</b>	GE: Something else	-	-
16	nl: lower secondary school basic profession-orientated learning path, year 4 (vmbo-b 4)	<b>15</b>	NL: Lower secondary school, basic professionorientated learning path, year 4 (VMBO-B 4)	-	-
17	nl: lower secondary school middle management-orientated learning path, year 4 (vmbo-k 4)	<b>16</b>	NL: Lower secondary school, middle management-orientated learning path, year 4 (VMBOk 4)	-	-



18	nl: lower secondary school mixed learning path, year 4 (vmbo-g 4)	<b>17</b>	NL: Lower secondary school, mixed learning path, year 4 (VMBO-g 4)	-	-
19	nl: lower secondary school theoretical learning path, year 4 (vmbo-t 4)	<b>18</b>	NL: Lower secondary school, theoretical learning path, year 4 (VMBO-t 4)	-	-
20	nl: intermediate secondary school, year 4 (havo 4)	<b>19</b>	NL: Intermediate secondary school, year 4 (HAVO 4)	<b>7</b>	NL: Intermediate secondary school, year 4 (HAVO 4)
21	nl: intermediate secondary school, year 5 (havo 5)	<b>20</b>	NL: Intermediate secondary school, year 5 (HAVO 5)	<b>8</b>	NL: Intermediate secondary school, year 5 (HAVO 5)
22	nl: upper secondary school, year 4 (vwo/gymnasium 4)	<b>21</b>	NL: Upper secondary school, year 4 (VWO/gymnasium 4)	<b>9</b>	NL: Upper secondary school, year 4 (VWO/gymnasium 4)
23	nl: upper secondary school, year 5 (vwo/gymnasium 5)	<b>22</b>	NL: Upper secondary school, year 5 (VWO/gymnasium 5)	<b>10</b>	NL: Upper secondary school, year 5 (VWO/gymnasium 5)
24	nl: vwo/gymnasium 6	-	-	<b>11</b>	NL: Upper secondary school, year 6 (VWO/gymnasium 6)
25	nl: lower tertiary school (mbo-opleiding)	<b>23</b>	NL: Lower tertiary school (MBO-opleiding)	<b>12</b>	NL: Lower tertiary school, dual programme (MBO-opleiding)
				<b>13</b>	NL: Lower tertiary school, fulltime programme (MBO-opleiding)
26	nl: apprenticeship	<b>24</b>	NL: Apprenticeship	-	-
27	nl: working	<b>25</b>	NL: Working	<b>14</b>	NL: Working
28	nl: something else	<b>26</b>	NL: Something else	<b>16</b>	NL: Something else
29	sw: upper secondary school, academic track"	<b>27</b>	SW: Upper secondary school, academic track	-	-
30	sw: upper secondary school, vocational track	<b>28</b>	SW: Upper secondary school, vocational track	-	-
31	sw: upper secondary school, provisional track	<b>29</b>	SW: Upper secondary school, provisional track	-	-

32	sw: i will not study but intend to work instead	<b>30</b>	SW: I will not study but intend to work instead	-	-
33	sw: something else	<b>31</b>	SW: Something else	-	-

**7.2 B – Documentation of Divergent Answers between Student and School Leaver**  
**Variables in Waves Three and Five**

<b>Content Sheet</b>	<b>H Variable</b>	<b>Wave</b>	<b>NEPS Student Var. (pTarget)</b>	<b>NEPS School Leaver Var. (pTargetCATI)</b>	<b>Number of Divergent Answers</b>
General Info.	H_sx	3	t700031	t700001	3
General Info.	H_dobm	3	t70004m	t70000m	9
General Info.	H_doby	3	t70004y	t70000y	12
Household Sit.	H_hhm1	3	t74305a	t743024	41
Household Sit.	H_hhm2	3	t74305c	t743025	111
Household Sit.	H_hhm5	3	t74305e	t743026	61
Household Sit.	H_hhm6	3	t74305f	t743027	54
Household Sit.	H_hhm8	3	t74305g	t743031	63
Household Sit.	H_hhm9	3	t741002	t741001	121
Migration Hist.	H_counSC	3	t400000_g1R	t405000	0
School Perf.	H_gm_ge	3	t724112	tf11227	0
School Perf.	H_gg_ge	3	t724111	tf11229	0
School Perf.	H_reps	3	t725020	t725000	0
Attitudes t.w. school	H_helpt	5	t22452i	t254052	0
Future Plans	H_infoedupar	3	tf0023c	t292404	0
Future Plans	H_infoeduor	3	tf0023d	t292404	0
Future Plans	H_infoeduint	3	tf0023h	t292407	0
Future Plans	H_infoeducoun	3	tf0023g	t292401	0
Future Plans	H_infoedutch	3	tf0023f	t292406	0
Future Plans	H_infoedumed	3	tf0023b	t292403	0
Future Plans	H_infoeducenter	3	tf0023a	t292402	0
Language	H_lpscS	5	t41030b	t41331b	0
Language	H_lpscW	5	t41030d	t41331d	0
Language	H_slanS	5	t41040b	t41341b	0
Language	H_slanW	5	t41040d	t41341d	0

## 7.3 C – Documentation Year-Overlap Linear Equating

Content Sheet	H Variable	CILS4EU Variable(s)	Wave CILS4EU	NEPS SC4 Variable(s)	Wave NEPS SC4	H_Wave	Same Wave?
Migration Hist.	H_stay	futsc1	1	t421010	2	x	<i>No</i>
School Perf.	H_spm	sspm	1	t66001a	1	1	Yes
School Perf.	H_spc1	sspsc	1	t66000c	1	1	Yes
Attitudes t.w. school	H_sesch	seff1	1	t66002c	1	1	Yes
Attitudes t.w. school	H_segrad	seff2, s_seff2	1	t66002b	1	1	Yes
Attitudes t.w. school	H_statp	stat1	1	t30535a, t30535b	2	x	<i>No</i>
Attitudes t.w. school	H_helpt	tenc1	1	t22452i, t254052, t254001	5	x	<i>No</i>
Attitudes t.w. school	H_suclo	sucpr1	2	t30035a	3	3	Yes
Attitudes t.w. school	H_sucint	sucpr2	2	t30035b	3	3	Yes
Attitudes t.w. school	H_sucup	sucpr3	2	t30035c	3	3	Yes
Future Plans	H_infoedupar	infofut2	2	tf0023c	3	3	Yes
Future Plans	H_infoeduor	infofut3, infofut4	2	tf0023d	3	3	Yes
Future Plans	H_infoedufr	infofut6	2	tf0023e	3	3	Yes
Future Plans	H_infoeduint	infofut7	2	tf0023h	3	3	Yes
Future Plans	H_infoeducoun	infofut8	2	tf0023g	3	3	Yes
Future Plans	H_infoedutch	infofut9	2	tf0023f	3	3	Yes
Future Plans	H_infoedumed	infofut10, infofut11	2	tf0023b	3	3	Yes

Future Plans	H_infoeducenter	infofut12	2	tf0023a	3	3	Yes
Future Plans	H_jobalt	impjob2	2	t66210n	3	3	Yes
Romantic Relat.	H_chilfu	futchi	1	t533010	2	x	<i>No</i>
Family Relat.	H_penc	penc1	1	t320403	1	1	Yes
Language	H_lpscS	lpoc1	1	t41030b, t41331b	1	1	Yes
Language	H_lpscC	lpoc2	1	t41030a	1	1	Yes
Language	H_lpscR	lpoc3	1	t41030c	1	1	Yes
Language	H_lpscW	lpoc4	1	t41030d, t41331d	1	1	Yes
Language	H_slanS	lpoc1	1	t41040b, t41341b	1	1	Yes
Language	H_slanC	lpoc2	1	t41040a	1	1	Yes
Language	H_slanR	lpoc3	1	t41040c	1	1	Yes
Language	H_slanW	lpoc4	1	t41040d, t41341d	1	1	Yes
Identity	H_idSC	idsc	1	t428050	2	x	<i>No</i>
Identity	H_idEC	idoc2	1	t428300	2	x	<i>No</i>
Religion	H_relign	rel2	2	t435000	3	3	Yes
Well-being	H_satl	sat1	1	t514001	1	1	Yes
Well-being	H_satsc	sat2	1	t514006	1	1	Yes
Well-being	H_sequal	sest1	1	t66003c	1	1	Yes
Health	H_genhea	genhea	2	t521000	3	3	Yes

## 8 References

- Alba, R. (1988). Cohorts and the dynamics of ethnic change. In Riley, M. W. (Ed.). *Social Change and the Life Course*. Newbury Park: Sage.
- Blossfeld, H.-P., Rossbach, H.-G. and Maurice, J. von (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft*, **14**.
- CILS4EU (2016). *Children of Immigrants Longitudinal Survey in Four European Countries. Codebook. Wave 1 – 2010/2011, v1.2.0*. Mannheim: Mannheim University.
- Dollmann, J., Jacob, K. and Kalter, F. (2014). Examining the diversity of youth in Europe: A Classification of Generations and Ethnic Origins Using CILS4EU Data (Technical Report). *MZES Arbeitspapiere - Working Papers*, **156**.
- Granda, P., Wolf, C. and Hadorn, R. (2010). Harmonizing Survey Data. In Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B.-E. and Smith, T. W. (Eds.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, New Jersey: Wiley, pp. 315–334.
- Gresch, C. and Kristen, C. (2011). Staatsbürgerschaft oder Migrationshintergrund? / Citizenship or Immigrant Background? *Zeitschrift für Soziologie*, **40**, 208–227.
- Hoffmeyer-Zlotnik, J. H. (2008). Harmonisation of demographic and socio-economic variables in cross-national survey research. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, **98**, 5–24.
- Jacobs, D., Swyngedouw, M., Hanquinet, L., Vandezande, V., Andersson, R., Horta, A. P. B., Berger, M., Diani, M., Ferrer, A. G., Giugni, M., Morariu, M., Pilati, K. and Statham, P. (2009). The challenge of measuring immigrant origin and immigration-related ethnicity in Europe. *Journal of International Migration and Integration*, **10**, 67–88.
- Kalter, F., Heath, A. F., Hewstone, M., Jonsson, J. O., Kalmijn, M., Kogan, I. and van Tubergen, F. (2016). Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) – Full version.
- Kolen, M. J. and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. New York, NY: Springer New York.
- LifBi (2021). *Study Overview: NEPS Starting Cohort 4 — Grade 9. School and Vocational Training — Educational Pathways of Students in Grade 9 and Higher. Waves 1 to 12*.
- Olczyk, M., Will, G. and Kristen, C. (2014). Immigrants in the NEPS: Identifying Generation Status and Group of Origin. *NEPS Working Paper*, **41a**.

- Singh, R. K. (2020/2021). *Adventures in ex-post harmonization*: GESIS - Leibniz-Institut für Sozialwissenschaften.
- Singh, R. K. (2020). Harmonizing Instruments with Equating. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, **6**, 11–18.
- Singh, R. K. (2021). Harmonizing Data in the Social Sciences with Equating. In Wolbring, T., Leitgöb, H. and Faulbaum, F. (Eds.). *Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter. Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute*. Wiesbaden: Springer VS, pp. 123–140.
- Skopek, J., Pink, S. and Bela, D. (2013). Starting Cohort 4: 9th Grade (SC4). SUF Version 1.1.0. Data Manual. *NEPS Research Data Paper. National Educational Panel Study (NEPS), University Bamberg*.
- StataCorp. (2021). *Stata Statistical Software: Release 17*: College Station, TX: StataCorp LLC.
- Wolf, C., Schneider, S. L., Behr, D. and Joye, D. (2016). Harmonizing survey questions between cultures and over time. In Wolf, C., Joye, D., Smith, T. and Fu, Y.-c. (Eds.). *The SAGE Handbook of Survey Methodology*. London: SAGE Publications Ltd, pp. 502–524.
- Würbach, A. and Abmann, C. (2023). The Composite Weight of CILS4NEPS: Joint Weighting of the CILS4EUSample and the Sample of Starting Cohort 4 of the German National Educational Panel Study (Wave1). Technical Report. *Leibniz Institute for Educational Trajectories*.