Data Observer

Jörg Dollmann*, Lena Arnold and Andreas Horr

CILS4NEPS – Unlocking Research Potential Through More Participants, More Schools and International Comparison: Harmonized Data for Research on Education, School-to-work Transition and Integration Processes for Adolescents in Germany, the Netherlands, Sweden and England

https://doi.org/10.1515/jbnst-2024-0016
Received January 12, 2024; accepted January 13, 2024

Abstract: The CILS4NEPS project combined and harmonized panel data from the Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) and Starting Cohort 4 of the German National Educational Panel Study (NEPS SC4). This unlocks additional research potential beyond the scope of both individual datasets by increasing sample sizes and enabling international comparisons of the NEPS data. Both, the combined dataset as well as additional material to reproduce and expand the harmonization are available to users. In this article, we will first introduce the individual datasets and their comparability and describe the steps of the harmonization process. Furthermore, we will present the comparable content between both surveys, the structure of the harmonized dataset, and weighting issues. Subsequently, we provide descriptive statistics, the analytical potential, and information on data access. Lastly, we will finish with an outlook on the continuing harmonization of CILS4EU and NEPS SC4.

Keywords: CILS4EU; data harmonization; education; immigrants; NEPS SC4; panel-data

***Corresponding author: Jörg Dollmann**, Mannheim Center for European Social Research MZES, University of Mannheim, Postfach, 68131 Mannheim, Germany; and Research Data Centre DeZIM.fdz, DeZIM Institute, Berlin, Germany, E-mail: joerg.dollmann@uni-mannheim.de
Lena Arnold, Mannheim Center for European Social Research MZES, University of Mannheim, Mannheim, Germany; and Graduate School of Social Sciences, University of Mannheim, Mannheim, Germany
Andreas Horr, Leibniz Institute for Educational Trajectories, Bamberg, Germany

JEL Classification: C8 (Data Collection and Data Estimation Methodology; Computer Programs)

1 Introduction

In recent years, the increasing focus on the topics of immigration and integration has led to several new data infrastructure projects that specifically address these issues (e.g. Dollmann et al. 2023). One alternative to collecting new data is to combine different datasets in order to open up new research opportunities. The aim of the CILS4NEPS harmonization project was exactly this: to unlock additional research potential by combining the two data sources "Children of Immigrants Longitudinal Survey in Four European Countries" (CILS4EU; Kalter et al. 2016) and "Starting Cohort 4 of the National Educational Panel Study (NEPS)" (NEPS SC4; Blossfeld and Roßbach 2019; Fuß, von Maurice, and Roßbach 2016) which would not be possible or only to a limited extent with the respective individual datasets. Data harmonization of already existing sources – called ex-post harmonization – is an important tool to enhance the analytical potentials of combined data over and above the individual datasets. In the social sciences, there is a growing number of similar harmonization projects (Dubrow and Tomescu-Dubrow 2016; Wysmułek, Tomescu-Dubrow, and Kwak 2021). As Doiron and colleagues outline, data harmonization increases "sample sizes that could not be obtained with individual studies, improves the generalizability of results, helps ensure the validity of comparative research, encourages more efficient secondary usage of existing data, and provides opportunities for collaborative and multi-centre research" (Doiron et al. 2013: 1).

The CILS4NEPS harmonization project addresses several of these issues. (1) In terms of sample sizes, a combination of the German part of CILS4EU and the Germany-only study NEPS SC4 is a useful enrichment for national analyses in the German context, as it allows for an increase in the number of cases for certain (ethnic or social) groups as well as for certain events (particularly transitions to certain forms of schooling or vocational education). This enables more differentiated analyses than with the two individual datasets on their own. (2) In terms of generalizability, obtaining results from two large-scale German studies instead of one increases the confidence to accept these results as valid. (3) In terms of comparative research, CILS4NEPS facilitates the usage of NEPS SC4 data for international comparisons, such as comparing the schooling and educational careers of young people in Germany with those in England, the Netherlands, or Sweden (the three countries besides Germany surveyed in CILS4EU). This article informs about the construction of CILS4NEPS, data access and promotes the future reuse of the data.

In the following, we will first introduce the individual datasets and their comparability, describe the steps of the harmonization process, comparable content, the structure of the harmonized dataset, and sample weighting, before providing descriptive statistics, the analytical potential, and information on data access. Lastly, we will finish with an outlook on the continuing harmonization of CILS4EU and NEPS SC4.

2 Data Sources and Their Comparability

2.1 CILS4EU

CILS4EU is an international longitudinal study whose aim is to investigate the integration of young people with and without an immigrant background in Germany, England, the Netherlands, and Sweden (Kalter et al. 2016; Kalter, Kogan, and Dollmann 2019). The target population of CILS4EU is the student body at regular schools, excluding special schools that attended grade 9 in the school year 2010/11. In order to achieve sufficient numbers of cases of students with an immigrant background, schools with a larger proportion of immigrants were oversampled (for more details on the general design of CILS4EU see Kalter, Kogan, and Dollmann 2019). In addition to this explicit stratification, the individual countries were additionally stratified implicitly; in Germany by state and type of school in order to take these characteristics into account proportionally to the population.

A total of about 19,000 youths were surveyed in the first wave (approximately 5,000 in Germany), with about half having an immigrant background. As part of the international funding, two additional waves of surveys were implemented between 2011 and 2013, with the second wave also taking place in the school context, where the fieldwork was administered by IEA-DPC. Participants in the third wave were surveyed outside the school context online, by mail, or by telephone. After the third wave, the German part of CILS4EU was included in the DFG's long-term program. In 2016, in the course of the sixth wave of the survey, a refresher sample was drawn in which adolescents or young adults with an immigration background were also disproportionately represented. Currently, data from nine waves (plus one wave on the COVID-19 pandemic) are available for Germany.

2.2 NEPS SC4

Besides CILS4EU, the harmonization project and this paper use data from the National Educational Panel Study (NEPS; see Blossfeld and Roßbach 2019). The NEPS

is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. It collects longitudinal data on skill development, educational processes, educational decisions, and educational returns in formal, nonformal, and informal contexts.

Starting in grade 9, the SC4 sub-study (NEPS Network 2023) examines pathways into and through upper secondary education as well as transitions into the vocational education system, higher education, and the labor market. The target population is the student body at regular schools and special schools that attended grade 9 in the 2010/11 school year. For this purpose, a stratified lump sample was drawn from regular schools and a sample of adolescents from special schools.

The adolescents were 14–15 years old at the time of the first survey, and survey data are available for a total of slightly more than 15,500 adolescents for the first wave, of whom about 37 percent have an immigrant background. Adolescents who continued to attend the selected schools were surveyed in the school context in subsequent waves. School leavers were followed outside the school context (mostly via CATI). The main survey at the schools was conducted as PAPI by IEA-DPC, and the CATI and CAWI surveys in the individual field were conducted by infas – Institute for Applied Social Sciences. Currently, data from thirteen waves (plus one wave on the COVID-19 pandemic) are available for SC4.

2.3 Comparability

The combination of the two datasets is appropriate because both CILS4EU and NEPS SC4 refer to a very similar target population (adolescents aged about 14–15 years) or, in the case of the German sub-study of CILS4EU, even to the very same population (adolescents in grade 9 in the school year 2010/11). CILS4EU implemented a very similar sampling approach at the school level as NEPS SC4, but in doing so, schools with high proportions of immigrants were drawn disproportionately often. The international sampling design of CILS4EU as well as the national sampling and fieldwork in the drawn schools in Germany was 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 the comparability and thus the useful combination of the two data sources. Importantly, the schools selected for the NEPS SC4 sample were excluded directly from the German CILS4EU sample. This means that the same students are not duplicated in the harmonized dataset.

3 Data Harmonization

In this section, we will describe the different steps of the harmonization procedure of the data from CILS4EU and NEPS SC4. Both, the CILS4EU and NEPS SC4 datasets contain different respondent groups in certain waves such as students, parents, or teachers. For the harmonization project described in this article, we included students as target persons only – but plan on extending the harmonization to further respondent groups (i.e. parents, teachers). In total, the first three waves of CILS4EU and the first six waves of NEPS SC4 were harmonized (for more information on how these different numbers of waves were matched see Section 5).

In the following, when we use the term harmonization, we explicitly mean ex-post harmonization of our data. In contrast to ex-ante harmonization of data, in which surveys are designed to be comparable before they are collected, ex-post harmonization refers to the harmonization of existing survey data into an integrated dataset (Granda, Wolf, and Hadorn 2010). In the case of CILS4NEPS harmonization ex-post harmonization is the strategy of choice because both studies have already collected data over a period of time. The goal of ex-post harmonization is to create a combined dataset with harmonized variables that come from different source datasets but are built on a common definition of the construct (Wolf et al. 2016). This combination of datasets can be done for both cross-national and national surveys.

Overall, no firmly established steps exist for ex-post harmonization of data, but usually the following steps are suggested (Granda, Wolf, and Hadorn 2010; Singh 2021; Hoffmeyer-Zlotnik 2008): (1) identify datasets to be combined, (2) identify similar questions in source questionnaires that offer potential for harmonization, (3) define target items that combine source variables into harmonized variables, (4) define and decide on harmonization strategies to create the target items, (5) map routines used during data harmonization to ensure replicability. We followed these steps for the harmonization of CILS4EU and NEPS SC4. In the next subsections, we will outline steps 2 to 5 and explain how each of these steps was implemented in the harmonization process.

3.1 Identification of Similar Content

It is crucial for ex-post harmonization that the harmonized variables in each source dataset measure similar constructs (Singh 2020/2021). Although variables do not need to be measured in the exact same way, a certain degree of similarity is necessary as combining variables that do not measure a similar construct in the different source datasets (i.e. concept mismatch) would introduce serious bias in the

harmonized dataset (Singh 2020/2021). To assess the similarity of variables, we considered both question-wording and answer categories.

3.2 Definition of Target Items

To define the target items, we constructed a coding table which offered a precise overview of how each source variable was coded. The coding table represents 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. We again assessed the comparability of source variables by checking whether the same constructs are measured with regard to question-wording, whether the construct is observable or manifest, and whether response categories are similar in their form and number. 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 (e.g. date of birth). Due to this accordance in construct similarity and identical or very similar response categories, a harmonization procedure could be carried out for this target item in the form of matching (and, if necessary, recoding) of the response categories. *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 answer categories (e.g. by lagging responses). For these items, a simple assignment of response categories would have led to bias in the dataset and analyses. Instead, we applied linear equating as a harmonization strategy, which we outline below. We recommend users of these items to validate them in their analyses. *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 (e.g. double- versus single-barreled questions). These variables were not harmonized, as this would have introduced serious bias in the harmonized dataset.

3.3 Decision on Harmonization Strategy

The decision on the harmonization strategy was based on the above-described classification of items: *matching* of response categories for unproblematic items; *linear equating* for more complicated items. For *matching*, we merged the answer categories of both source variables – depending on the coding of source variables, sometimes reverse coding response scales or combining several answer categories

into one category was necessary. In certain cases, multiple variables from CILS4EU or NEPS SC4 were combined into one harmonized variable (e.g. three CILS4EU and one NEPS SC4 variable into one harmonized variable).

Equating is a harmonization tool that addresses the problem that for latent constructs respondents' true answer scores are unobservable (Kolen and Brennan 2014; Singh 2020). This means that two respondents with the same true score could classify themselves in different scores on the same answer scale for two source items that intend to measure the same latent construct but slightly differ in their question wording, or vice versa (Kolen and Brennan 2014; Singh 2020/2021; Singh 2020). For equating to produce reliable results when used as a harmonization strategy, three prerequisites need to be fulfilled: First, equity property assumes that respondents have one true score on the latent item regardless of the survey – even though the observed scores may differ between surveys (Singh 2021). Second, the items to be harmonized should be measured in a similar way in both surveys. Third, to avoid temporal and spatial 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 fulfill these prerequisites in the CILS4EU and NEPS SC4 harmonization since both surveys encompass the same target population and we carefully assessed the comparability of the source items.

For the linear equating process, we selected CILS4EU variables as the target items – adapting the scale of each NEPS SC4 item to that of the respective CILS4EU target item. When applying linear equating, it is assumed that differences in the distribution of the observed scores are only due to differences between the measurement instruments. Therefore, we aligned the distribution of answers in the different surveys, as proposed by Singh (2020). An important assumption in this harmonization 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, the values of the source items are linearly transformed – meaning that the mean and standard deviation of the source and target item become equal (Kolen and Brennan 2014; Singh 2021). 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 harmonized score but scaled to the format of the target scale." (Singh 2021: 128). As equating requires the same target populations for its construction of a recoding table, we only obtained means and standard deviations from the German sample in CILS4EU. This recording table was then subsequently applied to the whole sample. We weighted the data during the linear equating process with the applicable individual weights to achieve a valid representation of the sample populations. The

following formula from Kolen and Brennan (2014) represents the linear transformation, where a variable x is transformed to the mean (μ) and standard deviation (σ) of variable y.

$$ly(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \quad (1)$$

$$slope = \frac{\sigma(Y)}{\sigma(X)}, \text{and } intercept = \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \quad (2)$$

3.4 Mapping of Routines

To map the routines of our harmonization process, we recorded each step of the process via tables or Stata scripts ("do-files") which are provided to users. In an 'overview table', we provide the full overview of all variables in the CILS4EU wave 1 to 3 and the respective similar NEPS SC4 items. The 'coding table' (see Section 3.2) contains information on the specific coding of each CILS4EU and matching NEPS SC4 item as well as detailed coding and instructions for the harmonized target item. For the linear equating process, we provide an Excel-document which includes the recoding tables and graphical representations of each linear equated item. Lastly, we provide all Stata do-files which include the technical construction of the harmonized dataset. Even though users can directly work with our finished harmonized dataset, this mapping allows for a full replication of this dataset.

4 Comparable Content

Both studies capture a wide variety of different constructs and concepts: sociodemographic information, information about educational plans, careers, transitions, social capital and friendships ties; attitudes and values on a variety of aspects of life and additional information stemming from cognitive and verbal achievement tests. However, the eventually comparable content after the harmonization steps outlined above is naturally much smaller. Nevertheless, we were able to harmonize over 100 items referring to diverse constructs and concepts, such as the household situation and composition, students' social background and immigration history, school performance, attitudes towards school, future plans, economic situation, current situation, romantic relationships, family relations, language, identity, religion, leisure time activities, well-being, and health. Out of the items covered by these constructs and concepts, about two thirds were harmonized with matching

techniques and about one third through linear equating. Information about the concrete items being harmonized and the underlying harmonization strategy can be found in the CILS4NEPS codebook under Section 2.2 (CILS4NEPS 2023) which is available at: https://www.neps-data.de/Data-Center/Data-and-Documentation/Start-Cohort-Grade-9/CILS4NEPS. As outlined above, the item classification presented there as either *unproblematic* or *more complicated* largely reflects the applied harmonization strategy of matching and linear equating.

5 Data Structure

We harmonized the first three waves of the CILS4EU with the first six waves of the NEPS SC4. This difference in the number of waves results from differences in the frequency of data collection between the two surveys. While CILS4EU collected data on a yearly basis in the first three waves, NEPS SC4 respondents were interviewed at different time points, depending on their status as students or school-leavers. Therefore, 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. While 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, some users might be interested in specific respondent groups (e.g. school-leavers), which were interviewed in wave 4 and 6. Due to this, we decided against the construction of one uniform harmonized wave variable that exactly matches the CILS4EU and NEPS SC4 waves. Instead, we constructed two wave indicators for the harmonized dataset. The first wave indicator ('H_wave') retains the original wave structure of the NEPS SC4 including six waves. Yet, as only three waves are available in CILS4EU for the respective time frame of data collection, waves 2, 4, and 6 from this wave indicator include NEPS SC4 waves only (for an overview of the waves and wave indicators see Table 1). Therefore, we constructed a second wave indicator ('H_wave2') which matches waves 1, 3, and 5 in the NEPS SC4 to waves 1, 2, and 3 in the CILS4EU dataset (we recommend this wave indicator for users conducting panel analyses with the harmonized dataset). Overall, only three waves can be used for panel analyses with the harmonized data. However, 'H_wave' would allow users to compare the two school-leaver waves in NEPS SC4 (wave 4 and 6) with wave 2 and wave 3 in CILS4EU, respectively.

The harmonized dataset of the different waves is provided in a long data format, with 'H_wave' or 'H_wave2', indicating the waves in which the variable was asked. The harmonized dataset includes the CILS4EU and the NEPS SC4 data as well as their harmonized variables. To provide a quick overview of 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. Harmonized variables contain the

Table 1: Tabulation of the two harmonized wave indicators 'H_wave2' and 'H_wave'.

	2010/11		2011/12		2012/13	
NEPS SC4	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
CILS4EU	Wave 1		Wave 2		Wave 3	
CILS4NEPS						
H_wave	Wave 1	Wave 2 (NEPS only)	Wave 3	Wave 4 (NEPS only)	Wave 5	Wave 6 (NEPS only)
H_wave2	Wave 1	–	Wave 2	–	Wave 3	–

prefix 'H_' both in their variable name and in their variable label. As CILS4EU is provided in a wide data format, it was converted into a long format prior to data harmonization. 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.

6 Weighting[1]

Both surveys, CILS4EU and NEPS SC4, provide survey weights to account for the stratified sampling approach in both surveys. The sampling strategy in both surveys was rather similar. In both studies, the primary sampling units were schools enrolling the target group (i.e. ninth graders). These schools were selected with probability proportional to size, with number of classes or number of students in schools being the measure of size. Explicit stratification of schools in CILS4EU was done by assigning schools to mutually exclusive groups according to their proportion of immigrant background students in schools. Schools with a higher share of immigrant students were subsequently oversampled. In NEPS SC4, explicit stratification was done along the line of school types. Here, schools for basic secondary education (Hauptschulen), Rudolf Steiner schools (Freie Waldorfschulen), comprehensive schools (Integrierte Gesamtschule) and special-needs schools had higher chances of being included in the sample (see Steinhauer and Zinn 2016). After the selection of schools, in both surveys, two school classes were selected, and all

1 For a detailed overview of the calculation of the composite weights in CILS4EU see Würbach and Aßmann (2023), which this section is strongly based on.

students in the classes were asked to participate in the survey. Based on these selection probabilities on school- and class-level, survey weights were constructed in both surveys.

Sample selection was done for both surveys by IEA-DPC. The institute avoided sampling the same schools in both surveys, as this would have tremendously decreased the probability of participation of schools. This was done by applying replacement rules for schools that were already sampled by one of the surveys and were selected again by the other sampling of the other surveys. Therefore, both samples can be seen as mutually exclusive and independent from each other.

To conduct analyses with the pooled sample, using the initial design weights of both surveys is not appropriate. Therefore, Ariane Würbach and Christian Aßmann[2] at the Leibniz Institute for Educational Trajectories provided a composite weight of grade 9 students in the school year 2010/2022. Details on this calculation can be found in Würbach and Aßmann (2023). In the harmonized dataset six harmonized weight variables are available for the German population (cf. Table 2):

Which weights specifically are used for the analyses lies in the data users' decision, however, it is advisable to use standardized weights in general. The harmonized weights include only German cases and are hence not available for respondents from the other three countries of CILS4EU. Technically, the harmonized weights would allow for pooled analyses across the four countries, with the harmonized w_t_CILS4NEPS_std being the most comparable to the CILS4EU house

Table 2: Weights available in CILS4NEPS.

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)

2 Ariane Würbach is Head of Research Unit Statistical Survey Methods, incl. Head of Working Unit Sampling, Weighting, and Imputation at the LIfBi; Christian Aßmann is Head of Department 3 – Research Data Center, Methods Development at the Leibniz Institute for Educational Trajectories.

weight ('houwgt'). However, we advise data users – at least in addition to pooled analyses – to estimate models separately per country and to compare coefficients. To do so, we constructed two additional weights for the harmonized 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)

7 Descriptive Statistics

In the following, we present descriptive statistics of basic sociodemographic variables of the harmonized data CILS4NEPS. Table 3 displays the composition of the CILS4EU-, NEPS SC4- and the resulting CILS4NEPS-sample with respect to sex, age, and immigrant status (differentiating between the majority population and then first, second, and third generation). As expected, we have an almost equal share of males and females in our sample, with slightly more males in England and Germany (also within NEPS SC4) and slightly more females in Sweden and the Netherlands. The age structure is also rather similar with respondents in all countries being at the age of 15 during the first interview in wave 1. Regarding the composition of the sample with respect to the immigrant and non-immigrant origin population, both German samples (CILS4EU Germany and NEPS SC4) are rather similar, which was expected given the similar sampling strategy. However, there are slightly more pronounced differences between the countries, reflecting different immigration histories of the countries and therefore also different compositions of their immigrant population.

8 Analytical Potential

As outlined in the introduction, one of the two main aims of CILS4NEPS was to enable an international comparison of NEPS SC4 data to data from England, the Netherlands, and Sweden. Such an international perspective will make it possible to answer questions about the impact of educational institutions and the educational system on disparities and similarities of educational pathways of different social and ethnic groups (e.g. Dollmann 2021; Triventi et al. 2020). However, schools are not only places where knowledge is acquired and educational transitions are made, but also

Table 3: Descriptive statistics for CILS4EU, NEPS SC4, and CILS4NEPS (weighted percentages in brackets).

	CILS4EU					NEPS SC4	CILS4NEPS	
	England	Sweden	Netherlands	Germany	Total	Germany	Germany	Total
Total	4,315	5,025	4,363	5,013	18,716	15,577	20,590	34,293
Sex								
Male	2,211	2,481	2,144	2,571	9,407	7,870	10,441	17,277
	(48.8)	(59.3)	(50.5)	(51.2)	(50.2)	(51.1)	(51.2)	–
Female	2,102	2,544	2,216	2,442	9,304	7,705	10,147	17,009
	(51.1)	(49.7)	(49.4)	(48.8)	(49.7)	(48.9)	(48.8)	–
Age	15.1	14.6	15.1	15.3	15.0	15.1	15.2	15.1
	(0.387)	(0.365)	(0.553)	(0.619)	(0.534)	(0.630)	(0.639)	–
Immigrant Status								
Majority	2,301	2,588	2,894	2,513	10,296	10,993	13,506	21,289
	(70.1)	(66.1)	(81.3)	(68.5)	(71.2)	(70.5)	(69.2)	–
1st Generation	610	653	296	535	2,094	1,000	1,535	3,094
	(9.93)	(8.1)	(3.5)	(6.2)	(6.9)	(6.4)	(6.4)	–
2nd	1,092	1,574	1,073	1,780	5,519	3,103	4,883	8,622
Generation	(14.7)	(21.2)	(13.2)	(22.6)	(18.0)	(20.0)	(21.1)	–
3rd Generation	311	210	100	185	806	471	656	1,277
	(5.3)	(4.6)	(2.1)	(3.4)	(3.8)	(3.1)	(3.2)	–

Note. Total numbers in the table are unweighted. The percentages are weighted – for CILS4EU cases with the CILS4EU weight (houwgt), for NEPS SC4 cases with the NEPS SC4 weight (w_t), for CILS4NEPS – Germany cases with the CILS4NEPS weight (w_t_CILS4NEPS_std). For age, the mean and weighted standard deviation is included instead of total numbers and weighted percentages. In the NEPS SC4 sample, age is calculated based on the time point of the wave indicator due to the missing of a specific interview date in the first wave. Respondents are classified as majority, 1st, 2nd, and 3rd generation as described in Dollmann, Jacob, and Kalter (2014) and Olczyk, Will, and Kristen (2014).

where friendship networks evolve and attitudes are shaped (e.g. Kroneberg, Kruse, and Wimmer 2021; Wuestenenk, van Tubergen, and Stark 2022). In this respect, the combined dataset offers new opportunities in a comparative perspective. Finally, the life of adolescents not only takes part in schools. Also, the question of the role of outside-school contexts on the development of young people can be addressed internationally with this new dataset, making it possible to compare the NEPS SC4 sample to the samples in England, the Netherlands and Sweden.

The second main aim of CILS4NEPS was to allow for more fine-grained analyses in the German context, differentiating between smaller ethnic and social groups. Furthermore, the increased sample size in the German context offers the possibility to analyze more specific transitions in the educational and vocational system. The following Table 4 demonstrate the increased research potential in this respect once CILS4NEPS data is used by providing information about educational trajectories and the respective number of cases following these trajectories for CILS4NEPS. As can be seen, some of these outcomes are rather rare (e.g. prolonging non-academic schooling in column 5). Differentiating between students with and without an immigrant background may be possible in the single datasets CILS4EU and NEPS SC4. However, more fine-grained analyses between different immigrant groups (e.g. Turkish or Polish origin) are simply not possible. Here, the number of cases in the combined dataset may offer increased possibilities for more in-depth analyses (cf. the part of the table under "Combined Sample: Total").

Table 4: Overview of the individual samples and the realized combined sample differentiated by education trajectory.

	Gymnasium	Vocational training	Academic upgrading	Prolonged non-acad. schooling	Vocational preparation/ other	All
NEPS SC4 Sample: Total	4,058	2,416	1,873	632	2,119	11,098
No immigrant background	3,291	1,916	1,340	373	1,466	8,386
Immigrant background	767	500	533	259	653	2,712
Generational status						
First generation	117	129	125	61	176	608
Second generation	646	368	403	197	469	2,083
Origin group						

Table 4: (continued)

	Gymnasium	Vocational training	Academic upgrading	Prolonged non-acad. schooling	Vocational preparation/ other	All
Turkey	113	104	127	87	163	594
Southern Europe	65	68	35	12	76	256
Former Yugoslavian Republic	54	57	37	27	68	243
Former Soviet Union/CEE	245	176	186	80	223	910
Northern and Western Europe	94	20	39	9	22	184
Other	189	71	107	43	94	504
Unknown	7	4	2	1	7	21
CILS4EU-DE	699	694	747	757	363	3,260
Sample: Total						
No immigrant background	446	432	394	342	190	1,804
Immigrant background	253	262	353	415	173	1,456
Generational status						
First generation	39	61	74	104	35	313
Second generation	214	200	278	310	137	1,139
Origin group						
Turkey	68	84	119	167	69	507
Southern Europe	12	39	39	31	16	137
Former Yugoslavian Republic	17	30	26	35	15	123
Former Soviet Union/CEE	73	60	77	90	34	334
Northern and Western Europe	16	9	7	8	8	48
Other	67	40	85	84	31	307
Unknown	0	0	0	0	0	0

Table 4: (continued)

	Gymnasium	Vocational training	Academic upgrading	Prolonged non-acad. schooling	Vocational preparation/ other	All
Combined Sample: Total	4,757	3,110	2,620	1,389	2,482	14,358
No immigrant background	3,737	2,348	1,734	715	1,656	1,0190
Immigrant background	1,020	762	886	674	826	4,168
Generational status						
First generation	156	190	199	165	211	921
Second generation	860	568	681	507	606	3,222
Origin group						
Turkey	181	188	246	254	232	1,101
Southern Europe	77	107	74	43	92	393
Former Yugoslavian Republic	71	87	63	62	83	366
Former Soviet Union/CEE	318	236	263	170	257	1,244
Northern and Western Europe	110	29	46	17	30	232
Other	256	111	192	127	125	811
Unknown	7	4	2	1	7	21

Note. Source: CILS4NEPS (version 1.0.0), with additional information from CILS4EU (version 3.3.0) and NEPS SC4 (version 13.0.0). Own calculations, balanced sample of respondents participating in (harmonized) Waves 1, 3, and 5. *Gymnasium*: Respondents in academic tracks in Wave 1 and Wave 5. *Vocational training*: Respondents in vocational training in Wave 5. *Academic upgrade*: Respondents in non-academic tracks in Waves 1 that were enrolled in academic tracks in Wave 5. *Prolonged non-acad. schooling*: Respondents in either non-academic tracks or vocational schools in Wave 5. *Vocational preparation/other*: Respondents in vocational preparation courses or other activities in Wave 5.

9 Data Access

The harmonized data product CILS4NEPS is available at the Research Data Center of the Leibniz Institute for Educational Trajectories (LIfBi) via remote access. Data access is granted to all scientific users who can demonstrate that they are interested in analyses of the harmonized dataset. Access exclusively to CILS4EU via the

Research Data Center cannot be granted. In this case, the application has to be made via the Gesis data archive.

As a first step, applicants apply for the reduced version of the CILS4EU data at the Gesis data archive for the social sciences. In this application, it needs to become clear why the research question can only be answered with the combined data product and not with one or both individual datasets. As a second step, applicants must then apply for the data of NEPS SC4. Again, a form with personal information, contact details and an outline of the planned research projects needs to be provided. To access the harmonized CILS4NEPS data, the NEPS Data Use Agreement has to be supplemented by an access authorization to the protected remote data processing of the LIfBi Research Data Center. For this RemoteNEPS extension, there is a separate form that also needs to be completed and signed.

The application forms are available via the following links:

https://www.gesis.org/en/institute/departments/data-services-for-the-social-sciences

https://www.neps-data.de/Data-Center

10 Summary and Outlook

The harmonized CILS4NEPS data provides unique opportunities for more fine-grained analyses on group- and trajectory-level which would not be possible with the individual CILS4EU and NEPS SC4 data alone. Furthermore, it enables researchers to relate the findings from the German NEPS SC4 to the results, particularly on (but not limited to) educational trajectories in three other European countries: England, the Netherlands and Sweden.

The procedures described in this data brief and the current version of the data product are the beginning of further efforts to harmonize both data sources. More precisely, we intend to combine subsequent waves above and beyond waves that were in the scope of the harmonization project so far. Seven further waves are currently available for the German part of CILS4EU (until wave 9 plus one Covid-19-wave) as well as for NEPS SC4 (until wave 13). For all of these waves, we investigate the harmonization potential and plan to implement the outcomes of our efforts in future data releases.

Additionally, we can extend our harmonization efforts to other target populations which were included in both CILS4EU and NEPS SC4. So far, we have only focused on students. In the future, we will also aim at harmonizing data from students' parents and their teachers. Finally, we will also have a closer look at different

achievement measures that were conducted in both surveys and evaluate the possibilities of bringing such measures together in a harmonized data product. However, we believe that the CILS4NEPS data product already offers interesting analytical possibilities already in its current state, especially in, but not limited to, the context of empirical educational research.

Acknowledgments: We are grateful to Markus Weißmann, Viktoria Kerzner, Hannah Soiné, Regine Schmidt, Florian Weber, Ariane Würbach, Christian Aßmann and Daniel Fuß for their help in the creation of the CILS4NEPS dataset and Markus Weißmann for his help in calculating and checking statistics for the present article. We are responsible for all possible remaining errors.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: This work was supported by KonsortSWD, Research Data Management Grant, Deutsche Forschungsgemeinschaft (KA 1602/8-3, and KO 3601/8-3).
Data availability: Not applicable.

References

Blossfeld, H.-P., and H.-G. Roßbach, eds. 2019. *Education as a Lifelong Process: The German National Educational Panel Study (NEPS)*, Edition ZfE, Vol. 2. Wiesbaden: Springer VS.

CILS4EU. 2014. *Children of Immigrants Longitudinal Survey in Four European Countries. Technical Report, Wave 1 2010/2011, v1.1.0*. Mannheim: Mannheim University.

CILS4NEPS. 2023. *Codebook: A Harmonised Dataset Based on CILS4EU and NEPS SC4 (CILS4NEPS), Version 1.0 with CILS4EU Waves 1-3 and NEPS SC4 Waves 1-6*. Mannheim: Mannheim University.

Doiron, D., P. Burton, Y. Marcon, A. Gaye, B. H. R. Wolffenbuttel, M. Perola, and R. P. Stolk. 2013. "Data Harmonization and Federated Analysis of Population-Based Studies: the BioSHaRE Project." *Emerging Themes in Epidemiology* 10 (12): 1–8.

Dollmann, J. 2021. "Ethnic Inequality in Choice-and Performance-driven Education Systems: A Longitudinal Study of Educational Choices in England, Germany, the Netherlands, and Sweden." *The British Journal of Sociology* 72 (4): 974–91.

Dollmann, J., K. Jacob, and F. Kalter. 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.

Dollmann, J., S. J. Mayer, A. Lietz, M. Siegel, and J. Köhler. 2023. "DeZIM. Panel. Data for Germany's Post-Migrant Society." *Jahrbücher für Nationalökonomie und Statistik* 243 (1): 93–108.

Dubrow, J. K., and I. Tomescu-Dubrow. 2016. "The Rise of Cross-National Survey Data Harmonization in the Social Sciences: Emergence of an Interdisciplinary Methodological Field." *Quality & Quantity* 50: 1449–67.

Fuß, D., J. von Maurice, and H.-G. Roßbach. 2016. "A Unique Research Data Infrastructure for Educational Research and beyond: The National Educational Panel Study." *Jahrbücher für Nationalökonomie und Statistik* 236 (4): 517–28.

Granda, P., C. Wolf, and R. Hadorn. 2010. "Harmonizing Survey Data." In *Survey Methods in Multinational, Multiregional, and Multicultural Context*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B-E. Pennell, and T.W. Smith, 315–34. New Jersey: Wiley.

Hoffmeyer-Zlotnik, J. H. P. 2008. "Harmonisation of Demographic and Socio-Economic Variables in Cross-National Survey Research." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 98 (1): 5–24.

Kalter, F., A. F. Heath, M. Hewstone, J. O. Jonsson, M. Kalmijn, I. Kogan, and F. van Tubergen. 2016. Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) – Full version. Data file for on-site use. Cologne: GESIS Data Archive, ZA5353 Data file Version 1.2.0, https://doi.org/10.4232/cils4eu.5353.1.2.0.

Kalter, F., I. Kogan, and J. Dollmann. 2019. "Studying Integration from Adolescence to Early Adulthood: Design, Content, and Research Potential of the CILS4EU-DE Data." *European Sociological Review* 35 (2): 280–97.

Kolen, M. J., and R. L. Brennan, eds. 2014. *Test Equating, Scaling, and Linking*. New York: Springer New York.

Kroneberg, C., H. Kruse, and A. Wimmer. 2021. "When Ethnicity and Gender Align: Classroom Composition, Friendship Segregation, and Collective Identities in European Schools." *European Sociological Review* 37 (6): 918–34.

NEPS Network. 2023. *National Educational Panel Study, Scientific Use File of Starting Cohort Grade 9. Version 13.0.0*. Bamberg: Leibniz Institute for Educational Trajectories (LIfBi).

Olczyk, M., G. Will, and C. Kristen. 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: Frankenstein's Creature – A Blog Post Series on Harmonizing Data from Different Surveys. GESIS – Leibniz – Institut für Sozialwissenschaften. https://blog.gesis.org/adventures-in-ex-post-harmonization-frankensteins-creature/ (accessed January 29, 2024).

Singh, R. K. 2020. "Harmonizing Instruments with Equating. Harmonization: Newsletter on Survey Data Harmonization." *Social Science* 6 (1): 11–8.

Singh, R. K. 2021. "Harmonizing Data in the Social Sciences with Equating." In *Schriftenreihe der ASI – Arbeitergemeinschaft Sozialwissenschaftlicher Institute: Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter*, edited by T. Wolbring, H. Leitgöb, and F. Faulbaum, 123–40. Wiesbaden: Springer VS.

Steinhauer, H.-W., and S. Zinn. 2016. "NEPS Technical Report for Weighting: Weighting the Sample of Starting Cohort 4 of the National Educational Panel Study (Wave 1 to 6)." *NEPS Survey Paper*. 2.

Triventi, M., J. Skopek, N. Kulic, S. Buchholz, and H.-P. Blossfeld. 2020. "Advantage 'finds its Way': How Privileged Families Exploit Opportunities in Different Systems of Secondary Education." *Sociology* 54 (2): 237–57.

Wolf, C., S. L. Schneider, D. Behr, and D. Joye. 2016. "Harmonizing Survey Questions between Cultures and over Time." In *Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T. W. Smith, and Y.-C. Fu, 502–24. London: SAGE Publication Ltd.

Wuestenenk, N., F. van Tubergen, and T. H. Stark. 2022. "Attitudes towards Homosexuality Among Ethnic Majority and Minority Adolescents in Western Europe: The Role of Ethnic Classroom Composition." *International Journal of Intercultural Relations* 88: 133–47.

Würbach, A., and C. Aßmann. 2023. *The Composite Weight of CILS4NEPS: Joint Weighting of the German CILS4EU Sample and the Sample of Starting Cohort 4 of the German National Educational Panel Study (Wave 1). Technical Report referring to* DOI:10.5157/CILS4NEPS:SUF:1.0. Bamberg: Leibniz Institute for Educational Trajectories (LIfBi).

Wysmułek, I., I. Tomescu-Dubrow, and J. Kwak. 2021. "Ex-post Harmonization of Cross-National Survey Data: Advantages in Methodological and Substantive Inquiries." *Quality & Quantity* 56: 1701–8.