DE GRUYTER OLDENBOURG Journal of Economics and Statistics 2016; 236(4): 517–528

Data Observer **Open Access**

Daniel Fuß*, Jutta von Maurice and Hans-Günther Roßbach

A Unique Research Data Infrastructure for Educational Research and Beyond: The National Educational Panel Study

DOI 10.1515/jbnst-2015-1021

Abstract: The article provides an insight into the conceptual and methodological framework as well as the research data infrastructure of the German National Educational Panel Study (NEPS). The NEPS study has been set up to build a profound empirical basis for the description and analysis of educational processes and competence development across the life span. Its large-scale database consists of longitudinal information from more than 60,000 target respondents – distributed over six different starting cohorts ranging from newborns to adults – and from relevant context persons such as parents or teachers. The complex multicohort sequence design schedules annual or even semiannual survey waves including a broad spectrum of competence assessments. All data are thoroughly prepared, documented, and disseminated free of charge in the form of regularly expanded Scientific Use Files. In addition to some background information about NEPS in general, this paper primarily focuses on issues of data collection, data structure, data availability, and the requirements for different types of data access. The number of more than 1,000 data users involved in over 700 research projects so far serves to highlight the potential of NEPS as a unique research data infrastructure for educational research and beyond.

1 Introduction

The National Educational Panel Study (NEPS) has been set up to find out more about how education is acquired, to understand how it impacts on individual biographies, and to describe and analyze the major educational processes and trajectories across the life span. It started in 2009, building a large-scale

***Corresponding author: Daniel Fuß,** LIfBi – Leibniz Institute for Educational Trajectories, Research Data Center (FDZ), Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: daniel.fuss@lifbi.de

Jutta von Maurice, Hans-Günther Roßbach, LIfBi – Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany

quantitative database with longitudinal data from more than 60,000 target respondents from six different starting cohorts supplemented with information from relevant context persons. This data infrastructure is provided free of charge to researchers all over the world in the form of thoroughly prepared and regularly expanded Scientific Use Files (SUF). The first NEPS data were made available in December 2011.

NEPS is carried out as an interdisciplinary activity that integrates theoretical and methodological approaches from sociology, psychology, educational sciences, economy, statistics, and other disciplines. As of February 2016, scientists from 18 universities and research institutes all over Germany contribute their special expertise to the NEPS Consortium.[1] From 2009 to 2013, NEPS was funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). In 2014, it was converted to the Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe, LIfBi) and became part of the Leibniz Association (Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz). LIfBi is a member of the Interdisciplinary Network of Infrastructure Facilities of the Leibniz Association (Interdisziplinärer Verbund wissenschaftlicher Infrastruktureinrichtungen, IVI) and the Leibniz Education Research Network (Leibniz-Forschungsverbund "Bildungspotenziale", LERN).

2 The conceptual framework of NEPS

Education is a highly interactive process. Every person has to make several *educational decisions* throughout his or her life (especially at a younger age these are usually made by or together with adults). The scope ranges from decisions regarding extrafamilial care, school enrollment, and the type of secondary school track, across decisions regarding vocational training, university studies, or withdrawing from the education system completely, to decisions

1 Members of this nationwide network of excellence are (in alphabetical order): Berlin Social Science Center (WZB), Centre for European Economic Research (ZEW) in Mannheim, German Centre for Research on Higher Education and Science Studies (DZHW) in Hannover, German Institute for International Education Research in Frankfurt (DIPF), Humboldt-Universität zu Berlin (HU), Ifo Institute–Leibniz Institute for Economic Research at the University of Munich, Institute for Employment Research in Nuremberg (IAB), Institute for School Development Research (IFS) at TU Dortmund University, IPN–Leibniz Institute for Science and Mathematics Education at Kiel University, Justus Liebig University Giessen, Leibniz Universität Hannover, Leipzig University, Ludwig-Maximilians-Universität in Munich (LMU), Universität Hamburg, University of Bamberg, University of Mannheim, University of Siegen, University of Tübingen.

regarding lifelong learning. Each of these choices might have a direct impact on the formal, nonformal, and informal *learning environments* that individuals enter or leave and that differ in structure, support, challenge, and orientation. Within these contexts cognitive and noncognitive *competencies* develop, irrespective of whether the environment is primarily designed for learning or not (e. g., schools vs. peers). The elements mentioned here are multiply linked, but of course, this is not a one-way street: Educational and career decisions depend on the competencies at hand, and participation in learning environments is significant for decision-making processes. Beyond this dynamic interaction, NEPS highlights monetary and nonmonetary *returns to education* (also competence can be seen as an "outcome" of education) as well as the role that any kind of *migration background* might play in this context.

At an overarching level, the framework of NEPS follows the general principles of life-course research and the perspective of life span developmental psychology. The five outlined dimensions are conceptualized as being prevalent during all life stages; however, particular attention is paid to eight phases of life or transitions:

- newborns and early childhood education,
- from Kindergarten to elementary school,
- from elementary school to lower secondary school,
- from lower to upper secondary school,
- upper Gymnasium level and transition to higher education, vocational training, or the labor market,
- from vocational training to the labor market,
- from higher education to the labor market, and
- adult education and lifelong learning.

A detailed description of the major theoretical dimensions and its central underlying assumptions is given in Blossfeld, Roßbach, and von Maurice (2011).

3 Design, samples, methods, and survey program of NEPS

In order to collect and provide data from early childhood to late adulthood within a reasonable period of time, a multicohort sequence design with six starting cohorts (SC) was implemented. The sampling took place between 2009 and 2012 either on an individual or an institutional basis. Individual samples were generated by registration offices; institutional samples were based on

school records or lists of subjects attending universities and universities of applied sciences. All cohorts are representative of Germany at the first wave. As Table 1 shows, starting sample sizes vary between 3,007 children for the Kindergarten cohort (SC2) and 17,911 persons for the first-year student cohort (SC5). In the younger cohorts (SC1–SC4), context persons, such as parents and educational staff, were additionally included in the survey in order to gain a deeper insight into the respective learning environments. At some limited points of the design, additional target respondents became integrated into the NEPS design. Sample replenishments or refreshments were realized when the Kindergarten children (SC2) reached school age in 2012 (+6,342 respondents), when the fifth graders cohort (SC3) reached Grade 7 in 2012 (+2,205 respondents), and when the adult cohort (SC6) was complemented in 2011/2012 by a sample of individuals born between 1955 and 1986 (+5,208 respondents).

All participants are surveyed regularly, usually at least once a year. Different modes are applied as there are paper-and-pencil questionnaires (PAPI), computer-based telephone and personal interviews (CATI/CAPI), as well as online surveys (CAWI). The competence assessments are conducted as paper-and-pencil tests or as computer-based tests with laptops or tablets either in group or individual settings. In the youngest cohort (SC1), videotaped observations are used too.

As already pointed out, following the development of domain-specific competencies across the life span is a key objective of NEPS. The focus is on reading and listening, mathematics, and the natural sciences. These measures are supplemented by the assessment of domain-general cognitive functioning, metacompetencies such as metacognition or ICT literacy, and – in selected starting cohorts – stage-specific competence tests such as orthography in elementary school or English in secondary school (for an overview see Artelt/Weinert/Carstensen 2013). Related constructs such as interests and motivation, personality and self-regulation, as well as social competencies are also included in the survey program.

The conceptual dimension of learning environments is covered by several questions on structure, excitation, and support features of formal (e. g., schools), nonformal (e. g., corporate trainings), and informal (e. g., peer groups) learning contexts. Within the realm of educational decisions, indicators of social and cultural capital are surveyed across the starting cohorts as well as constructs referring to rational choice theories. Broad information is available on the migration background of NEPS respondents, ranging from origin, nationality, and language use to identity, religion, and assimilation. A special emphasis is placed on Turkish migrants and ethnic German repatriates. Returns to education are measured in the form of economic benefits such as income, wealth, and

Table 1: Sampling characteristics and participation rates by NEPS starting cohorts.

Starting cohort	Sample strategy (field start of wave 1)	Context persons in the survey	Panel sample of targets at Wave 1[a] (% of gross sample)	Augmentation of target sample in later waves	Completed waves as of December 2015	Panel sample of targets at last completed wave[a] (panel stability)
Early childhood (SC1)	Individual (08/2012)	Parents, childminders, day-care workers	3,439 (40.5 %)	–	4	3,022 (87.9 %)
Kindergarten (SC2)	Institution-based (01/2011)	Parents, educators, administrators, class teachers, school principals	3,007 (56.2 %)	6,342	5	6,002[d] (86.8 %)
Grade 5 (SC3)	Institution-based (11/2010)	Parents, class teachers, math teachers, German teachers, school principals	6,112 (52.9 %)	2,205	6	7,325 (88.1 %)
Grade 9 (SC4)	Institution-based (11/2010)		16,425 (61.1 %)	–	8	13,272 (80.8 %)
1st-Y. Students (SC5)	Institution-based (11/2010)	–	17,911 (27.2 %)[b]	–	9	14,529 (81.1 %)
Adults (SC6)	Individual (02/2009)	–	13,576 (50.3 %)[c]	5,208	6	11,541 (61.4 %)

[a]The number of realized cases might be lower due to soft refusals (temporary dropouts); the number of cases in SUF might be lower due to ex post request by participants for data deletion.

[b]A larger number of 21,097 respondents (32.0 % of gross sample) submitted panel consent and valid contact information; the panel sample in SC5 was additionally restricted to participation in Wave 1.

[c]The sample of SC6 consists of two subsamples – that is, participants of the former study "Arbeiten und Lernen im Wandel" (ALWA; $N = 4{,}926$ respondents) and newly sampled participants ($N = 6{,}615$ respondents).

[d]The entire cohort consists of 9,337 cases. Of the initial Kindergarten sample ($N = 3{,}007$ respondents) 11 cases droped out, 576 cases transitioned to a NEPS primary school, and the remaining 2,420 cases are to be surveyed again in Wave 6. Thus, the number of students to be surveyed from Wave 3 to Wave 5 consists of 6,918 cases.

career or class mobility, but also in the form of noneconomic benefits such as health, subjective well-being, and social or political participation. The NEPS survey program is rounded off with rich social background data and detailed biographical information on schooling and education, vocational training and employment, course participation, partnership and family.

4 NEPS data structure and availability of scientific use files

Making available the data collected by NEPS for research requires a thorough preparation and editing process. Before any SUF can be released to the scientific community, a set of complex steps including data cleaning and editing, anonymization, coding, enrichment, documentation, and registration has to be conducted. The main responsibility for dealing with these issues lies with the Research Data Center at LIfBi (RDC LIfBi). Since 2012, the RDC LIfBi is an accredited member of the research data infrastructure of the German Data Forum (Rat für Sozial- und WirtschaftsDaten, RatSWD) and, thus, complies with the criteria for providing access to high-quality microdata for scientific purposes.

The data portfolio of NEPS contains six ongoing longitudinal studies with several panel waves and two finalized school-reform studies with two or three cross-sectional survey waves. Given the annual or semiannual data collection sweeps along the six starting cohorts, the challenge of ensuring a maximum of data usability becomes obvious. First of all, the demand for a consistent and clearly defined data structure is met by the provision of integrated longitudinal data sets for all cohorts. Every time the data of a new survey wave are prepared, the existing SUF for the respective starting cohort is extended by these data and released as a new version. Second, data are provided with homogeneous labels and in a similar structure across waves and cohorts. Each SUF consists of multiple data files representing different types of content and different data formats:

- *Cohort profile* is a standard data set in long format that we create for the users' convenience and recommend as a starting point for any analyses with the NEPS data. On the one hand, it includes all relevant identifiers for matching the information from different data files. On the other hand, the cohort profile provides an overview of the participation status, information availability, and sample assignment for each wave and each target respondent.
- *Survey files* basically include information in the way it was collected from the interviews, broken down by the type of respondent (target person,

parent, educator/teacher) and the reference of information (institution). These integrated data sets are structured in long format where each data row corresponds to the information given by a respondent at a certain wave.
- *Competence file* is a single data set in wide format that contains scored item variables for all conducted competence assessments, but also sum and mean values of correct test answers, as well as weighted likelihood estimators and corresponding standard errors from item response theory (IRT) scaling procedures.
- *Grouping files* provide information about the formal learning environments of the target respondents such as the Kindergarten group, the class, or a certain course cluster. These data are collected from the respective educators and teachers. In most cases, information about one group such as class size or class composition applies to a number of target respondents. The records in this file are in long format where one row represents one group at one time of measurement, bearing in mind that groups are not persistent over time.
- *Spell files* are generated data sets to ease the analysis of the rich life-course information collected retrospectively and prospectively over several waves. Episodes of general education and vocational training, employment and unemployment, civil and military service, and parental leave, as well as entity-related incidents such as partners, children, and siblings are carefully edited and stored in separate files. The data are in long format with each row representing one duration spell or one entity spell.
- *Additionally generated files* represent either condensed biographical information that is derived from the more detailed spell data sets (e. g., education) or data that might be necessary to be considered in analyses (e. g., weights).
- *Paradata files* are part of all SUFs as they contain important methodological information from the data collection process, also including interviewer evaluations of the survey or test situation and some interviewer characteristics. The data are also structured in long format.
- *Regional data files* are an extra feature of the NEPS portfolio. A broad variety of small-scaled regional indicators – ranging from street level to municipality – are acquired from commercial suppliers and linked to the respondents' survey data. Access to these files is restricted due to their sensitive content.

The consistent structure in connection with common editing conventions facilitates data handling on the part of the researchers. A further improvement on the quality and utility of NEPS data has been achieved by adding numerous derived variables. By default, standard scales measuring socioeconomic status and

occupational prestige (e. g., Magnitude-Prestige scale, SIOPS-88/SIOPS-08, ISEI-88/ISEI-08, EGP classes) are delivered as well as indicators for levels of educational attainment (e. g., CASMIN, ISCED-97). Likewise, standard codes for occupations and vocational education (e. g., KldB-88/KldB-2010, ISCO-88/ISCO-08), for business sectors and industries, for adult and further education courses, and for fields of study in higher education are given whenever appropriate. With regard to spatial information, several variables containing codes for administrative regions according to official district numbers and the NUTS hierarchy of regional clusters are available.[2] Last but not least, some other variables are generated addressing more specific issues such as indicators of personality traits or migration background.

The richness and complexity of NEPS data require intense efforts to prepare and compile them as an easy-to-use data product. Especially the process of (re) coding open entries – that is, the assignment of a numerical code from a selected category scheme or classification to the string information – and the integration of longitudinal information including the transformation of life history data into spell data sets are extensive and time-consuming. As soon as the data are prepared and checked, the respective SUF is made available to the scientific community. The general aim is to publish the data no later than 18 months after the fieldwork for the respective survey wave has finished.

Before any SUF is released, a unique and persistent digital object identifier – a so-called DOI – is assigned.[3] This identifier allows researchers to cite NEPS data directly in their publications in an easy and precise way. It further ensures the traceability of the research process; a demand that has become more and more important in the context of good scientific practice.

5 Data security and access to NEPS data

Facilitating good scientific practice by ensuring the highest possible extent of information represents one central concern of the NEPS endeavor; another basic

2 Since 2015, the use of the Federal State variable ("Bundeslandzugehörigkeit") in conjunction with NEPS data collected in connection with schools or higher education institutions is also permitted for defined purposes. In other cohorts, the use of the Federal State variable is not restricted.

3 The assignment is carried out at da|ra, the German registration agency for social and economic data. The DOI code indicates the relevant NEPS starting cohort and the data version of the SUF. It also refers to a landing page at the NEPS web portal that informs about basic metadata and ways of data access.

requirement is to guarantee a maximum of confidentiality protection for the – partly underage – survey respondents and their individual microdata. In order to comply with strict national and international standards, a bundle of data security measures and a flexible data access concept have been established. According to three hierarchical levels of information sensitivity, the NEPS data are available either by secure *download* from the website, or by a remote-access technology via virtual desktop called *RemoteNEPS*, or at *on-site* workstations at the RDC LIfBi in Bamberg. Data files that are available "on-site" are anonymized to the weakest degree and, thus, provide more sensitive information than the data files in the "remote-access" environment. The "download" version of data files features the highest level of anonymization.[4] For instance, almost all information from the students' or parents' questionnaires is disseminated through all three access modes, whereas school-related data are available in the controlled environments of *RemoteNEPS* and on-site only, and more fine-grained regional indicators are accessible exclusively on-site in Bamberg.[5]

Any access to NEPS data requires a signed data use agreement, and only persons who are affiliated with a scientific institution are eligible to apply for it. Interested researchers are requested to provide a brief description of their intended project and to specify the expected duration of data usage as well as further participants in the project. The relevant forms in German and English language can be retrieved from the NEPS website. Once the scientific purpose of the application has been approved, all available SUF data can be used free of charge. The data packages are provided in Stata and SPSS format with both English and German labels.

4 All data are of course de facto anonymous. Any identifiable information is coarsened or cut off in order to minimize the risk of statistical disclosure. String variables relating to openly asked questions are thoroughly checked.

5 The remote data access option requires an individual authentication through the researcher's keystroke biometrics. It provides a safe and powerful handling of sensitive NEPS data in an online research environment equipped with common statistical software packages and tools. The on-site data access option is bound to a stay as guest researcher at LIfBi. Data work takes place within a controlled physical environment – that is, all input and output devices are locked down and the computer is not connected to the Internet or any other local area networks. In order to prevent any copying or removing of sensitive data from the LIfBi server, each output from remote or on-site data usage passes through a review before being provided to the researcher.

6 Documentation and NEPS user services

Each NEPS SUF is accompanied by comprehensive documentation materials that inform researchers about how the data have been collected and prepared and how to make use of them. On the one hand, these materials consist of classically produced and regularly updated papers such as data manuals, method reports, release notes, descriptions of competence tests, working papers, and technical reports (e. g., on weights, anonymization, regional indicators). On the other hand, we rely on a structured and integrated approach to metadata management. It rests upon a relational SQL database that enables efficient storing and linking the mass of metadata on NEPS studies, instruments, data sets, variables, answer schemes, filtering instructions, etc. in a systematic and consistent fashion across waves and starting cohorts. The central maintenance of NEPS metadata in one big database allows for a dynamic documentation. It ensures a high documentation utility because every correction and extension becomes effective in a synchronous and consistent way in all related materials. Bilingual codebooks, survey instruments, and data set labels are the most important concomitant documentation materials, but also the interactive exploration tool *NEPSplorer* obtains its contents from this database. This software tool offers a full text search through all recorded metadata of survey instruments. For each item, related information on question phrases, corresponding variables, answer categories, interview instructions, etc. can be browsed and – according to individual requirements – stored in a user's personal watch list. Further auxiliary tools are, for instance, provided in the form of specific Stata syntax packages (*NEPStools*). All materials are available on the NEPS website.

The above-mentioned criteria of the German Data Forum for the establishment and administration of a research data infrastructure not only posit that a detailed documentation of data is issued, but also that far-reaching services and individual advice for data users are offered. In the case of NEPS, this requirement is of particular significance, as the multifaceted survey design creates a complex database with many data files. In order to facilitate well-informed and proper data usage, the RDC LIfBi holds more than a dozen 2-day training courses and data workshops per year and maintains an e-mail and a phone hotline for individualized counseling.[6]

6 An overall objective of NEPS training courses is the sensitization of researchers to issues of privacy and data protection. For that reason, course participation is obligatory for data users who want to enroll in the biometric authentication system for gaining access to the secure *RemoteNEPS* and on-site environment. The e-mail address for inquiries at the RDC LIfBi is: *fdz@lifbi.de*.

7 NEPS data usage

Although NEPS is a relatively young panel study, the number of researchers who use this database for their analyses has steadily risen over the last couple of years. By the end of January 2016 – about 5 years after the first SUF release – more than 700 research projects with a total of more than 1,050 data users involved had been registered. About 10 % of these projects are conducted at research institutions abroad, for instance, in the United States, the United Kingdom, Russia, India, Australia, Italy, or Switzerland. A somewhat closer look at current NEPS data usage reveals various disciplines and a broad spectrum of research questions ranging from learning processes and competence development to issues with regard to labor market, migration, social mobility, or family, but also to methodological topics and educational reporting.[7]

8 Summary and outlook

The combination of six starting cohorts, different types of information, multiple informants, the broad range of covered topics, and the longitudinal design with detailed biographical modules and repeated measures makes the NEPS database a highly attractive and unique source of empirical research in the field of education and beyond. Although there is a clear focus on competence development and educational processes – taking into account relevant life-course-specific learning environments as well as issues of social inequality – the range of possible analyses is by no means limited to this. While the potential of the NEPS data is going to increase steadily with respect to both addressing new research questions and employing more elaborated techniques of analysis, the data also bear the potential of being utilized as a benchmark when referring to them as a representative comparison group or as a reference for the case of Germany in international comparisons.

The most significant challenges for the future of NEPS are the constant improvement of survey and test instruments (e. g., increasing use of online surveys and computer-based testing modes with branched or adaptive testing), the further development of the survey design (e. g., prolonged interview or test intervals, restart of cohorts), and a continuing increase in user-friendliness and

7 A sorted list of all NEPS-based projects including further information can be found on the NEPS website: https://www.neps-data.de/en-us/datacenter/researchprojects.aspx.

data usability (e. g., broadened data portfolio by additionally generated variables and linked information from other data sources).

All relevant information about NEPS and its database are provided in our regularly published *LIfBi Data Newsletter* and at our website: *https://www.neps-data.de/*.

References

Artelt, C., S. Weinert, C.H. Carstensen (Eds.) (2013), Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS). Journal for Educational Research Online (JERO) 5: 5–240.

Blossfeld, H.-P., H.-G. Roßbach, J. von Maurice (Eds.) (2011), Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]. Zeitschrift für Erziehungswissenschaft 14: 5–330.