Psychological Test and Assessment Modeling, Volume 59, 2017 (1), 35-54

Large-scale assessments: potentials and challenges in longitudinal designs

Jutta von Maurice[1], *Sabine Zinn*[2] *& Ilka Wolter*[2]

Abstract

The article elaborates on the benefits and challenges of implementing a longitudinal design into large-scale assessments in educational research. Thus, the focus lies on educational trajectories and competence development within the population under study as well as the relevant processes behind them. Taking the starting cohort of ninth graders of the German National Educational Panel Study as an example, more detailed information is given on the aspect of sampling and selectivity at the start of a longitudinal large-scale study, as well as tracking and bias while keeping a panel running. Concerning instrumentation, challenges and methods connected with the measurement of competence development and the valid recording of individual biographies are discussed.

Keywords: longitudinal research, competence development, educational trajectories sampling, selectivity

[1] *Correspondence concerning this article should be addressed to:* Jutta von Maurice, PhD, Leibniz Institute for Educational Trajectories, Executive Director of Research, Wilhelmsplatz 3, 96047 Bamberg, Germany, WP3/02.41; email: jutta.von-maurice@lifbi.de

[2] Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany

1. Introduction

Large-scale assessments are of utmost importance in educational research. Studies like the Programme for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS), or the international Progress in Reading Literacy Study (PIRLS) lead to fundamental knowledge gains concerning students' competencies at different ages, competence distributions in total as well as in subgroups, and covariations of students' competencies with variables at the family level (social or migration background) or the school level (teacher and school characteristics, school type, class composition). Among other things, PISA allows us to analyze the competence level of 15-year-olds in great detail with a special emphasis on their learning environments. By means of the PISA data of 2012 (OECD, 2014) it can be shown that within the OECD countries 23% of all students show very low mathematical competencies (Level 1 or below) whereas 3% reach very high mathematical scores (Level 6). The corresponding analyses conducted show correlations between mathematical competencies and gender, as well as between migration background and social status. However, these results do not allow us to derive any explanations of how and at what age different competence levels of adolescents develop and which learning environments (not only formal, but also non-formal and informal) foster or compensate the effects of given background aspects. Furthermore, the effects of a given competence level on mid-term or long-term educational and vocational trajectories and life-course development cannot be studied by means of the PISA data (because they are cross-sectional). Concretely, PISA does not allow us to answer any of the following questions: How will competencies develop in the future? How successful are adolescents with low, median, or high competence levels in achieving a school degree, finding an apprenticeship position, or entering higher education and the labor market? What conditions can promote the educational and vocational processes even in adolescents with Level-1 (or below) competencies or hamper these processes even in the most competent Level-6 students?

Because studies such as PISA can only take a snapshot of the competencies of students of a certain age group, questions about developments, about supportive or destructive environments, about processes (e.g., in aspiration setting) cannot be dealt with. Instead of referring to a single-time-point measurement, questions such as those above can only be worked on by using longitudinal data (also called panel data). This paper aims to highlight the potential of large-scale assessments following a longitudinal design. For this purpose, the National Educational Panel Study (NEPS; cf., Blossfeld, Roßbach, & von Maurice, 2011) with a special emphasis on the educational pathways of ninth graders is used as an illustrating example of a general elaboration on the specific characteristic of such a design. The NEPS is based on the principles of life-course research (Elder, Johnson, & Crosnoe, 2003; see also Elder & Giele, 2009) as well as the perspective of life-span developmental psychology (Baltes, 1990; Baltes, Reese, & Lipsitt, 1980). Against this theoretical background, the clear focus is on *educational trajectories* as well as *competence development* and on the *relevant processes* behind them.

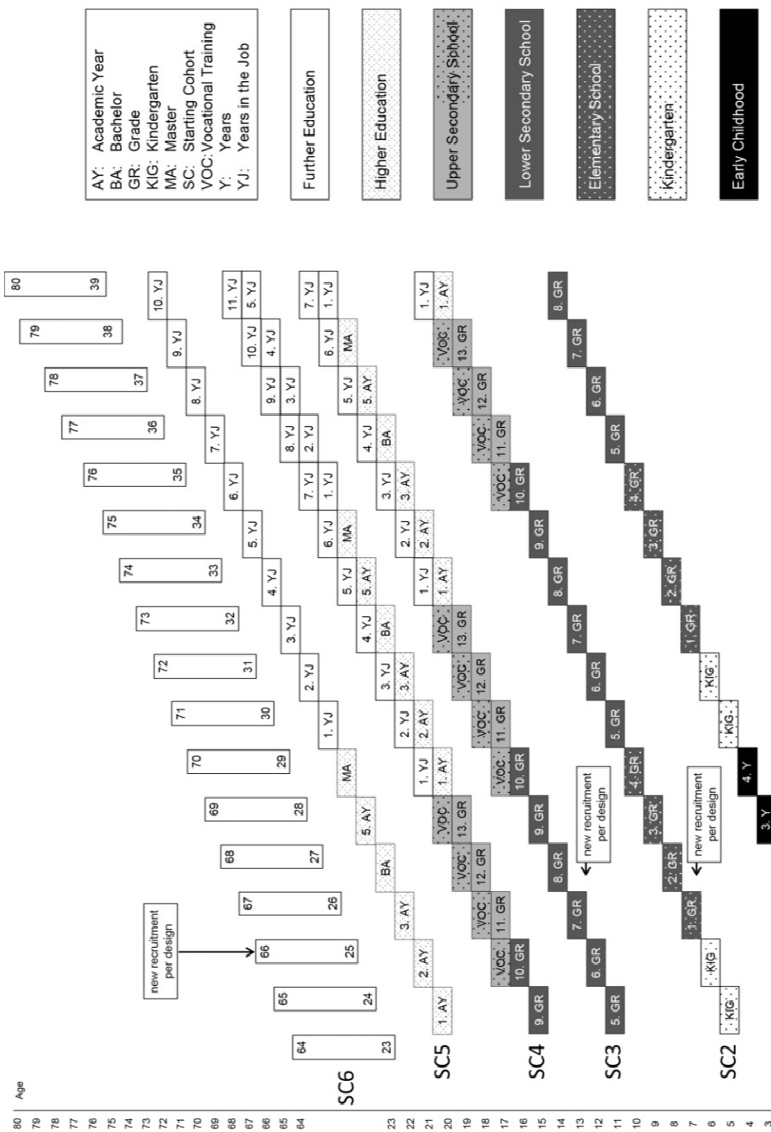

Figure 1:
Multicohort sequence design of the National Educational Panel Study with six starting cohorts (SC).

The NEPS was built up with six different starting cohorts (SC) ranging – at Measurement Point 1 – from early childhood through Kindergarten and school to higher education and adults aged 23 to 64. Figure 1 shows the design of the multicohort sequence design implemented in the study.

In order to follow educational processes and competence development in these six starting cohorts over time, appropriate sample sizes had to take into account the expected variations of multiple life trajectories. All cohorts were sampled at an individual or institution-based level – all are representative of Germany. Choosing a longitudinal instead of a cross-sectional design (such as the one in PISA) allows us to gain information about life-course development, which requires considering the variety of individual characteristics as well as the learning and living conditions at a single point in time but also the development of those characteristics and conditions over time. Therefore, longitudinal studies need large sample sizes in order to cover a multiplicity of life courses and must also anticipate the (expected) panel attrition (process) already from the beginning (see also Sections 2 and 3). The six starting cohorts contain about 60,000 target persons. All these individuals are followed over long time spans, resulting in a *longitudinal large-scale design*.

After recruitment (for sampling and selectivity see also Section 2), all participants are followed along their individual life courses (for tracking and bias see also Section 3). Regular measurement consists of questionnaires and tests of various competence domains, such as general cognitive capacities or metacompetencies (or precursors of domain-specific competencies in the youngest cohorts). For reasons of test settings in the distinct starting cohorts, the application of questionnaires and tests is provided in different modes ranging from one-to-one personal interviewing situations across group-based classroom settings to individually applied online mode. Questionnaires are mostly administered in a computer-based way in order to integrate complex filter or check routines to ensure reliable answers (e.g., open responses of birth countries should match existent countries). In some cases, also paper-based surveys are administered – for example, in group-based assessments in school classes. Besides the targets (i.e., participants of our cohorts), also context persons are involved in the investigation up until the time when targets leave secondary school. First of all, parents are regularly interviewed as they are not only a "marker" of social and ethnic background but also form a long-lasting learning environment for the developing individuals under study. Second, people involved in extrafamilial care or formal learning environments are regularly interviewed. It is only through the combination of these data sources that an in-depth analysis of the developmental processes within different learning environments becomes possible (for more detailed information, see Blossfeld et al., 2011).

In this way, traditional and nontraditional educational and vocational careers of panel participants, as well as the formal, nonformal, and informal environments they enter and leave, can be documented and analyzed. Furthermore, the age, period, and cohort effects of social inequality and of migration background can be monitored. By administering regular competence tests to all participants, we cannot only paint a precise picture of the competence level at single time points but also describe competence development. The longitudinal data collected facilitate a wide variety of analyses that would not be possible

with cross-sectional data. For example, longitudinal data allow for causal inference (see, e.g., Arjas & Parner, 2004; Allison, 2005) and event history analysis (cf., Allison, 1984; Singer & Willett, 2003).

In this paper, we use the starting cohort 4 (SC4) of the NEPS as an illustrating example. At the onset of the study, the participants of SC4 had just entered Grade 9 as part of the lower secondary school system in Germany (for a description of the German educational system please refer to Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, 2015). A particularity of the SC4 sample is that, at panel start, a main subsample was close to passing from lower school tracks in secondary education (i.e., Hauptschule [school for basic secondary education]) over to vocational training or the so-called transition system [option supporting the transition into the dual system]. Other subsamples of the SC4 study comprise students who continue schooling in Grade 10 or further – for example, to achieve university entrance qualifications (i.e., Gymnasium [type of school leading to upper secondary education and general qualification for university entrance]). Due to the variety of possible educational pathways after completing lower secondary education, the SC4 sample allows very detailed and comparative analyses concerning competence development and educational processes of adolescents in distinct situations and contexts. Therefore, the SC4 data of the NEPS enable researchers to gain some deeper insight into educational trajectories and competence development according to formal schooling in Germany and furthermore to investigate, for example, the impact of social disparities or educational decisions on competencies in this highly important transition phase.

As mentioned before, students in Grade 9 are schooled in different educational tracks. This is why the SC4 study started with a larger sample size than for example the younger cohorts in NEPS. This age cohort is of very high relevance for empirical educational research because it also facilitates international compatibility with other large-scale assessments such as PISA, which examines a comparable group of 15-year-old students. In the year 2016, all SC4 participants will have left school and started (or maybe even finished) their vocational education, university studies, or will have entered the labor market. The transition from the end of lower secondary education into vocational education, higher education, and employment is one of the particular interests of the SC4 and also one valuable characteristic of a longitudinal large-scale assessment such as the NEPS.

In the following, Section 2 describes the challenges of constructing a representative sample for a longitudinal large-scale study. In this context, the problem of selective attrition and wave nonresponse is also addressed. Thereafter, Section 3 discusses tracking issues. In connection with this, the problem of selection bias is elaborated and possible ways to counteract it are presented. Section 4 deals with the problem of how to accurately measure competence development across the life course and record biographical data. Finally, Section 5 summarizes the challenges and the potential of longitudinal large-scale studies such as the NEPS by giving a conclusion.

2. Sampling and selectivity

It is the research aim that determines the population from which to draw the sample and the sampling design. For example, the objective of the panel study SC4 of the NEPS is to inquire into the situation of ninth graders in Germany, their competence acquisition, as well as their educational and adolescent pathways. This objective necessitates a large amount of longitudinal data reflecting the personal, societal, and contextual aspects significantly related to the educational processes taking place in this stage of life. Concretely, in SC4 around 16,500 adolescents are surveyed in their institutional and individual contexts over time. In addition, a variety of context data such as institutional and parental information are collected. In total, in the SC4 study (until Wave 6) 8,813 parents have been surveyed to date together with 2,230 teachers in 542 schools.[3]

A longitudinal large-scale study places high demands on establishing the sample. First of all, the sample has to be a representative portrayal of the target population, as only a representative sample allows us to derive unbiased estimates of the particular population phenomenon of interest. To give an example, the target persons of the SC4 study are all adolescents attending secondary school in ninth grade in Germany in the school year 2010/2011 (Aßmann et al., 2011). Until that point, the requirements on sampling do not differ from cross-sectional designs. However, in a longitudinal design we have to consider the fact that ninth graders are a very heterogeneous population with a great variety of possible trajectories (see Figure 2 for illustration). Thus, to facilitate statistical inference, large samples have to be built comprising a significant number of persons for each

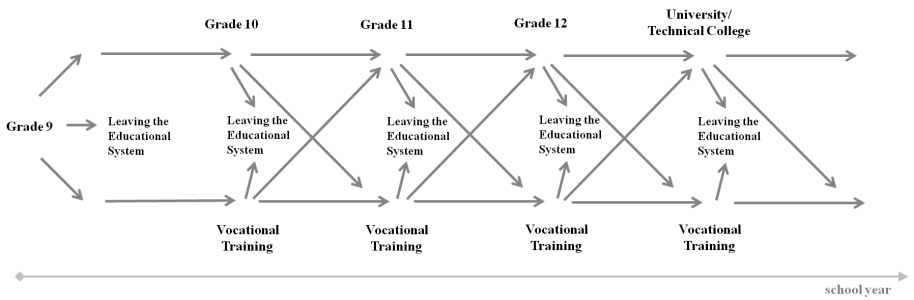

Figure 2:
Possible educational trajectories of the Starting Cohort 4: Grade 9 of the National Educational Panel Study (simplified illustration).

[3] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:6.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. For further details see Blossfeld, Roßbach, & von Maurice (2011).

possible trajectory. A common strategy to handle rare subpopulations or subpopulations that are hard to access (e.g., persons with a low level of education) as well as to ensure a sufficient number of rare trajectories is disproportionate stratified sampling (Kalton, 2009a, 2009b). Under this sampling scheme, the rare subpopulation to be studied forms an explicit stratum on its own. A meaningful sample size is obtained by assigning to this stratum a higher sampling fraction than compared to all of the other explicit strata formed. Accordingly, the SC4 sample is designed to consist of six explicit strata reflecting the distinct school types implemented in Germany.[4] In order to reach a meaningful number of educational trajectories (allowing for statistical inference) also of those students expected following unusual pathways, students in *Hauptschulen*, *Freie Waldorfschulen* and *Integrierte Gesamtschulen* are oversampled.

A stratified two-stage cluster sampling approach (see, e.g., Valliant, Dever, & Kreuter, 2013) was applied to gain access to the target population. At the first stage, a sample of all officially recognized and state-approved schools providing schooling to ninth-grade students was drawn. Thereafter, at the second stage, school classes were selected from those sampled schools that were willing to participate in the survey. Regarding the availability and accessibility of school data in Germany and the special structure of the population this approach has several advantages over, for example, simple random sampling. First, in Germany, no national database exists recording all ninth graders and the schools in which they are enrolled. However, school level data are available, on request, from the regional Statistical Offices of the 16 German Federal States (KMK, 2015). Thus, we established a sampling frame by requesting from each of the Statistical Offices a list of the (officially recognized and state approved) schools offering schooling to ninth graders. On that basis, we draw first a school sample and then a sample of ninth graders. The mode of data collection constitutes a further reason for multistage sampling. In the SC4, we used group tests in schools because they are easier to administer and more cost-efficient than individual tests. Finally, multistage sampling allows us to assess effects caused by clustering – such as differences in competence development across schools and classes (e.g., effects of class composition on competence development).

The participation in our study is voluntary. Hence, the problem of total nonresponse and wave nonresponse among the sampled units also has to be dealt with when constructing the initial sample. Total nonresponse refers to units that entirely refuse to participate in the study or fail to be surveyed in any wave. Wave nonresponse takes place when a sampled unit participates in some but not in all waves. That is, as opposed to cross-sectional studies, wave nonresponse in longitudinal studies is a critical issue to be addressed when determining the initial sample size. We use information from pilot studies that have

[4] The first stratum comprises *Gymnasien* [type of school leading to upper secondary education and certificate for university entrance], the second stratum consists of *Hauptschulen* [school for basic secondary education], the third stratum is made up of *Realschulen* [intermediate secondary school], and the fourth stratum covers *Integrierte Gesamtschulen* [integrated comprehensive schools] and *Freie Waldorfschulen* [Waldorf Schools]. The fifth stratum consists of *Schulen mit mehreren Bildungswegen* [schools with several courses of education], and the sixth stratum comprises *Förderschulen* [special schools].

specifically been conducted for that purpose to assess the response rates over the consecutive stages of the survey, and thus to derive the expected amount of nonresponse.

Overall, a total of 14,900 ninth-grade students were targeted to participate in the survey (Aßmann et al., 2011). In order to meet this objective, 1,749 schools were contacted in the 16 Federal States and asked for their participation in the NEPS.[5] In sum, 649 schools gave their participation consent, yielding an initial gross sample of 26,868 ninth-grade students. In the end, 16,425 students agreed to take part in the SC4 study (more details on the sampling design of the SC4 study and accordant numbers are given in Steinhauer, Aßmann, Zinn, Goßmann, & Rässler, 2015).

In the data collection endeavor, cross-sectional studies have to invest in engaging people to participate for only a single time point. To this end, information and persuasion have to be neatly intertwined with a correct handling of data protection aspects. These same issues also apply to longitudinal studies. However, in addition to this, further aspects related to the time horizon of a panel study have to be taken under consideration: Engaging participants in a longitudinal study is much more complex because the investment in time and the amount of information collected over the years are much greater. When recruiting the NEPS sample which includes ninth graders, parents, teachers as well as school heads, the idea of the entire panel had to be outlined. For this purpose, elaborate information material adapted to these different groups (flyers, brochures, and letters) as well as target-appropriate monetary and nonmonetary incentives were used. Moreover – in close collaboration with data collection institutes especially engaged for this purpose – the NEPS has invested a lot of resources in interviewer training so that interviewers are able to set up a comfortable interview situation and build a trustful relationship with the panel respondents. To run a panel study targeted at ninth graders, personal information such as names, addresses, and phone numbers of students and parents are needed over the whole duration of the study. Whereas longitudinal studies within the school context could even be conducted without the names and addresses of the students, a longitudinal study that also follows students as they move on to other schools, into apprenticeship and employment cannot be conducted without any concrete name and contact information.

The sampling units of SC4 are subject to unequal selection probabilities.[6] Disregarding this aspect in statistical analysis may lead to biased population estimates and misleading research conclusions. A common way to compensate for unequal selection probabilities is the use of design weights (see e.g., Särndal, Swensson, & Wretman, 2003; Pfeffermann & Rao, 2009). Design weights are typically defined as the inverse of inclusion

[5] To cope with the refusals of schools, for each sampled school a set of replacements schools (similar concerning their sampling and stratification characteristics) was defined. If a school refused participation, the replacement schools were successively contacted and asked for participation. Thus, when designing the SC4 sample at the school level, nonresponse was not assumed. In contrast, on the student level a total nonresponse rate of approximately 45% had been assumed based on the outcomes of preceding pilot studies.

[6] This is opposed to a direct sampling approach that uses simple random sampling. Here, every sampling unit has the same probability to be selected. The result is a self-weighted sample. Hence, in statistical analysis the sampling design is of no importance.

probabilities of the sampled units. The representativeness of the SC4 sample is impaired by selective refusals and total nonresponse.[7] To give an example, we find that students who speak a language other than German at home have a significantly lower participation propensity than their counterparts who only speak German at home. Likewise, male students have a lower participation propensity than female students (cf. Steinhauer et al., 2015). To this end, nonresponse adjustments of design weights can be used in statistical inference (Rosenbaum & Rubin, 1983, Kalton & Flores-Cervantes, 2003). Furthermore, (nonresponse adjusted) design weights are often additionally calibrated to correct for possible coverage errors and to reduce standard errors (see e.g., Valliant et al., 2013). All the steps used to compute the survey weights of the initial sample of a longitudinal study are similar to the weighting procedures applied in cross-sectional studies. However, counteracting and compensating for wave nonresponse is a much more challenging issue, and topic of the next section.

3. Tracking and bias in longitudinal studies

Usually, panel studies face the highest fraction of refusals and nonrespondents at the initial wave. Nonresponse rates of successive waves are commonly lower (see e.g., Lepkowski & Couper, 2002). This is exactly the pattern in the SC4 study where the initial overall fraction of nonresponse was around 40%, contrasted by an average wave nonresponse rate of about 15%. However, keeping response rates high over the course of the panel requires large investments in panel care. If contacted for the first time, the participation propensity of (future) panel members only relies on an illustration of the idea and the aims of the study. However, in the long run this cannot hold. After only very few waves, panel participants will begin to claim results. Often these expectations conflict with the time needed for data management and scientific analyses. Additionally, in a longitudinal setting a change in certain (often legal) conditions throughout the process of a particular study might cause problems. For example, students in Germany reaching the age of 14 have to sign their own letter of consent. Also, changing family constellations (e.g., new partnerships of participating mothers) or a change in the legal guardians' status may require new consent forms (or the exclusion of certain items from the respective parent interview). Last but not least, it is a challenging task to keep the questions and items interesting and motivating for the participants. The latter aspect is especially crucial when adolescents are constantly asked to report their lack of success in finding an appropriate apprenticeship place or job position over several successive panel waves. Consequently, panel management teams have to deal very carefully with the question of whether the materials and incentives that are provided to the target persons are appropriate to all interesting subsamples.

Typically, subgroups of special interest will fail to participate at some point, such as persons with a migration background in the SC4.[8] Here, within five years 6% of the

[7] Units that are sampled but never participate in a study are referred to as unit nonresponse.

[8] A person is assumed to have a migration background if at least two of the student's grandparents were born abroad (Olczyk, Will, & Kristen, 2014).

persons in the initial sample were lost due to attrition nonresponse – compared to 4% of persons without migration background. In other words, selection bias is evident and likely to increase over time. Among survey designers and data analysts, the biasing effect of panel attrition is a much noted issue (e.g., Magnusson & Bergmann, 1990; Alderman, Behrman, Kohler, Maluccio, & Watkins, 2001; Young, Powers, & Bell, 2006; Lynn, 2009), and several means to counteract this problem have been proposed (see e.g., Ribisl, Walton, Mowbray, Luke, Davidson, & Bootsmiller, 1996; Coen, Patrick, & Shern, 1996; Lynn, 2009).

One major source of panel attrition occurs because panel members can no longer be reached since they have moved or changed contact information, or both. A variety of approaches exist to track down panel members. For example, e-mails, postcards, or newsletters are sent to the last known contact address. This way, panel members are not only reminded of the study and thus encouraged to continue participating in the survey. On the other hand, the validity of contact information can also be proved and – if needed – new address information can be requested from the local registry offices. As an alternative or an addition, incentives might be given to limit attrition. In the SC4, a mix of strategies encouraging panel participation and counterbalancing bias effects is employed: Postcards are sent, and at every wave monetary incentives are paid out to all the participants. Concretely, panel members were grouped into two categories of respondents in order to work against different participation probabilities in different subpopulations: low risk and high risk of nonreponse. The latter group was defined as students from *Förderschule* and *Hauptschule* leaving the educational system. This "high-risk group" was paid a higher monetary incentive than the low-risk group (e.g., 15 Euro for the low-risk vs. 30 Euro for the high-risk group in the 2015 telephone interview). Furthermore, a lottery with attractive prizes (e.g., a car) was conducted for those who did not pass on to the *Gymnasium* after they had left the school context. Additionally, this group was also given information material with little text as well as an online link to the NEPS website where a film targeting this particular population explained the main aims of the NEPS.

A further method for enhancing response rates is the usage of mixed-mode surveys (see e.g., Dillman, Smyth, & Christian, 2014). Here, panel members are surveyed using more than one mode of data collection. Possible options are face-to-face interviews (in group settings or one-to-one settings), telephone surveys, mail surveys, and internet surveys. The success of a mixed-mode survey depends on several factors. Especially, the context and the target population are crucial. For example, the SC4 population comprises individuals in vocational training or transition system while others have entered *Gymnasium*. The first group is very mobile and hard to access. Thus, to ensure high response rates its members were asked to attend in a telephone survey. The latter group had been surveyed in groups in schools. This strategy has led to response rates of around 71% in the first group and response rates of around 86% in the second group. Also the survey instruments can be adapted to the participants in order to lower respondent burden. Within the NEPS, competence tests are adjusted to the respondents' competence level by branched testing in subsequent waves in order to avoid frustrating experiences during test taking. Finally, test instructions might be adapted to the specific needs of the relevant sample –

for example, students with special educational needs (e.g., Nusser, Carstensen, & Artelt, 2015).

Despite all the strategies applied to increase response rates, in the SC4 study wave nonresponse is still commonplace and selection bias occurs. That is, some survey units are more prone to wave nonresponse than others – for example, highly mobile individuals. For instance, attrition mostly occurs among adolescents who leave the school system after Grades 9 or 10 and enter the vocational track or the transition system (Steinhauer & Zinn, 2016). Reasons for this are outdated contact information and changes in living conditions (see also Ristau, Meixner, Sixt, & von Maurice, 2015). A further issue in this regard is item nonresponse. In detail, even if panel members participate in the survey wave, they might refuse to respond to certain answers. A well-known phenomenon in this respect is item nonresponse in connection with highly sensible questions. Usually, nonresponse is nonignorable in the sense that the answers (are expected to) differ between respondents and nonrespondents. Disregarding this aspect in statistical analysis will probably yield biased estimates and thus invalid research conclusions.

Different approaches exist to cope with item and wave nonresponse. One approach is the imputation of missing values (see e.g., Rubin, 2004; Twisk & de Vente, 2002). To this end, available responses from across all waves are used for each panel member. Another related method of dealing with item and wave nonresponse is full information maximum likelihood estimation (cf. Allison, 2001, 2003). Here, all available data are used to estimate a specific model. Alternatively response/nonresponse weighting adjustments might be used. Here, for each panel member or subgroup of special interest (e.g., adolescents who have participated in two successive competence tests) participation propensities are estimated by means of – for example – response propensity modeling. These propensities might enter a weighted analysis and facilitate the computation of unbiased estimates. Finally, information collected in subsequent waves is used to fill gaps caused by wave nonresponse (for dependent interviewing see Section 4). Despite all strategies in maintaining the panel, in handling statistical data and in instrumentation, a panel attrition (and corresponding bias) eventually is a severe limitation of all longitudinal studies.

4. Measuring development across the life course

Measuring development across the life course within the NEPS is connected to competence measurement as well as documenting individual life trajectories (cf. Blossfeld et al., 2011). In order to focus on a life span perspective such as provided in NEPS various supplemental aspects must be acknowledged when measuring competencies accordingly. On the one hand, the measurement instruments need to be adapted to the distinct cognitive developmental stages across the life course. On the other hand, and even more importantly within this longitudinal project, the instruments have to be designed to capture lifelong learning processes. This is opposed to only focusing on curriculum-based school subjects where usually shorter life periods are of interest (cf. Artelt, Weinert, & Carstensen, 2013; Weinert, Artelt, Prenzel, Senkbeil, Ehmke, & Carstensen, 2011). In a longitudinal design such as the one implemented in the NEPS, cohort invariant measures and

constructs that are coherently measured across the life course are differentiated. The cohort invariant measures function as baseline measures and are investigated either by domain-general or domain-specific competencies. In the NEPS, cohort invariant measures cover, for example, basic cognitive functioning (domain-general; cf. Haberkorn & Pohl, 2013) or reading speed (domain-specific; cf., Zimmermann, Gehrer, Artelt, & Weinert, 2012). Commonly, these constructs are only measured once at the beginning of a panel study, and are mostly used as control variables in analyzing the competence development of participants. Furthermore, stage-specific competence areas are measured, such as phonological awareness in younger children or job-related knowledge in adults. These skills are relevant in certain educational phases and reflect on the adaptation to stage-specific knowledge or precursor competencies. Over the course of the panel, they are not measured regularly or repeatedly.

Other domain-specific competencies as well as some meta-competencies are measured coherently across the life span. The coherent measuring and modeling of competence areas across the various educational stages, even outside the years of formal schooling, is the utmost challenge in longitudinal large-scale assessments such as the NEPS. Clearly, only those competence areas that are repeatedly and coherently measured will provide the possibility to analyze participants' actual growth in these competencies and thus their lifelong learning processes.

To comply with this requirement, for example, the competence of the NEPS school cohorts is repeatedly measured at intervals of 2 years. Precisely, in the SC4 surveys of Wave 1 and Wave 2, when students were in Grade 9, mathematical competence (Duchhardt & Gerdes, 2013), ICT literacy (Senkbeil & Ihme, 2012), scientific literacy (Schöps & Saß, 2013) and reading competence (Haberkorn, Pohl, Hardt, & Wiegand, 2012) were measured. A repeated measurement of those domains took place in Grade 11 for scientific literacy and in Grade 12 for mathematical competence, reading competence, and ICT literacy.

Basing test development within a longitudinal design on frameworks is a prerequisite to ensure a coherent measurement of domain-specific competencies over the life span. This aspect is even more important when competence measures are intended over diverse educational contexts from preschool throughout the ability-tracked school system in Germany and even up to education in different stages of adulthood respectively. To this end, test development of – for example, domain-specific competencies in NEPS – also follows specific frameworks to ensure that the construct being measured will be comparable even for the repeated measurement with different test instruments at different age groups. Domains that are based on such frameworks in NEPS are mathematical competence (Neumann, Duchhardt, Grüßing, Heinze, Knopp, & Ehmke, 2013), reading competence (Gehrer, Zimmermann, Artelt, & Weinert, 2013), scientific literacy (Hahn, Schöps, Rönnebeck, Martensen, Hansen, Saß, Dalehefte, & Prenzel, 2013), and also the meta-competencies such as ICT literacy (Senkbeil, Ihme, & Wittwer, 2013) and (declarative) metacognition (Händel, Artelt, & Weinert, 2013). The coherent assessment of competence in the different educational stages allows us to model competence acquisition over time in a methodologically sound way.

In SC4, the necessity of frameworks becomes obvious when considering the transition from secondary school to vocational training or higher school tracks (i.e., *Gymnasium*). It is the transition to educational stages outside formal schooling that highlight in particular the need to consider a competence definition in the longitudinal assessment of competencies, which does not focus only on curriculum-based skills. The NEPS frameworks also integrate core aspects of domain-specific competencies, regardless of a person's current life stage. This is especially relevant when accounting for the fact that competence development and educational trajectories continue after participants have left school. Constantly changing learning environments are an additional challenge in developing competence tests. For example, in the framework of measuring ICT literacy process components are combined with certain software applications to measure a person's ability in this domain (Senkbeil et al., 2013). Test development is therefore a challenging endeavor as these abilities are rapidly changing according to technological progress. For test development this implies the necessity to keep up with standards and modifications in the daily routines of the application. Another example for developing new standards and considering alternating demands for test development is the consideration of educational reforms, such as the German spelling reform in language competence tests.

Whenever multiple tests are applied to the same subject, as it is typically found in longitudinal designs, there is a need to link test scales to use and interpret scores on different test instruments. An obstacle in the repeated application of the same items is that items might have different levels of difficulty for respondents of different ages. Concretely, solving a particular item might be easier for a 16-year-old than for a 14-year-old student. However, linking allows us to associate test scores on a common scale and to compare the measures from subsequent time points. Linking procedures that will allow for measures to be used interchangeably and to make the outcome scores equated or comparable are widely conducted within large-scale assessment studies (cf., Carstensen, 2009; von Davier, Carstensen, & von Davier, 2008). According to Dorans (2000), there are some requirements to equating test scores: First and foremost, equal constructs with equal reliability are needed. The item response theory (IRT) provides approaches for linking procedures to use IRT parameters on a common scale and thus facilitates comparability of sequential measurements (cf., von Davier & von Davier, 2010). Adapted to the requirements and challenges regarding data collection within the IRT scaling approach of NEPS (cf., Pohl & Carstensen, 2013) a linking procedure is applied by interlinking the repeated competence measures over time by mean/mean linking the data of a subsequent measurement point to the according initial scale (Fischer, Gnambs, Rohm, & Carstensen, 2016). Here, similar reliability of the instruments is assumed and different item difficulties are allowed. The latter enables us to compare targets that are part of different subpopulations whose composition will possibly change over time. In NEPS, this situation arises, for example, when panel members grow older and transfer to different educational contexts. In other words, linking facilitates the comparison of competence attainment of the persons surveyed in NEPS over the life course.

There are various ways to conduct a link study for test instruments in a longitudinal large-scale study such as the NEPS (cf., Fischer, Rohm, Gnambs, & Carstensen, 2016). In some competence domains, this means that a few items (i.e., questions or tasks in a test instru-

ment) are repeatedly administered over subsequent studies (i.e., anchor-item design). In NEPS, such a design is, for example, applied for the domain mathematics (cf., Fischer et al., 2016). A slightly different approach is the anchor-group design: This approach is used for domains in which a repeated administration of the same test items within the same group is not possible, for example when higher memory effects are to be expected (cf., Fischer et al., 2016). In NEPS, reading competence is such a domain. The possibility that a participant will remember a specific text and the corresponding items prevents the entire test instrument or even a single text of an instrument from being administered repeatedly. One way to overcome this problem is to define an independent sample of test recipients (i.e., participants who are not included in NEPS cohorts) and to administer all test instruments that have to be linked in this sample (for a detailed description see Fischer, et al., 2016). Again, this kind of linking study allows for the transformation of competence measures to a common scale. In SC4, until now the domain-specific competence areas mathematical competence and reading competence were measured in Grade 9 and Grade 12. By linking both measures, the corresponding test scores become comparable on a common scale and, thus, competence development becomes visible.

Besides measuring competence development over the life course, the study of lifelong learning processes also demands that developments in other areas, such as professional life or family life, are also considered. Admittedly, there are some cross-sectional studies already collecting such data. However, the common means for this purpose are longitudinal studies. Here, the first part of the individual biography is usually collected retrospectively. That is, at Wave 1 panel members are asked to report life-course events from the past. Then, starting from Wave 2 biographic data are collected prospectively, that is, the individual trajectories are updated in every following wave. Over several waves, this way of processing yields highly detailed biographical information. Collecting such data needs very careful instrumentation. Because respondents might have difficulties to remember exact event dates or might even forget about events, self-reported biographies are very error-prone. Thus, recorded events and transition times might be inconsistent, resulting in unexplained gaps in biographies (e.g., a gap between two school spells) or nonconvincing parallel episodes (e.g., reporting a school episode and a parallel full-time job position). Temporal inconsistencies might be clarified either by applying consistency checks within the data collection process or by later data cleaning procedures. The first method is preferred to the second because the respondents can be actively involved in the data cleaning process. Within NEPS, this step is implemented in the so-called data revision module (see for more detail Ruland, Drasch, Künster, Matthes, & Steinwede, 2016). Nevertheless, of course, careful data editing is also needed after data collection to improve data quality (see for more detail Bela, 2016).

In longitudinal large-scale studies such as the NEPS, the biographical data of participants are updated at every wave (e.g., all educational steps taken after leaving school are recorded at every wave in SC4). Therefore, it is not sufficient to ask only about current situations. Instead, it is necessary to collect all relevant biographic spells since the last interview; this may mean since the last measurement wave or – in the case of wave nonresponse – since the last realized measurement. Taking SC4 as an example, the reconstruction of the educational trajectories is quite simple as long as the participants can be

found in the collaborating (NEPS) schools. However, as soon as the adolescents leave this school context to enter vocational training or the transition system, higher education or the labor market, highly heterogeneous pathways may emerge that are often associated with short educational episodes.

An established method of collecting reliable and consistent survey data is proactive dependent interviewing (for more detail, see Trahms, Matthes, & Ruland, 2016). Here, information from previous waves is integrated within the computer-administered instruments used for interviewing in the form of so-called preloads. That way, information such as the date of the last interview or information concerning the last educational status can be played back to the target person. Clearly, for the respondent it is much easier to answer the question whether a particular status is still applicable or if something has changed since a particular fixed date than the general question of asking whether "something changed since the last interview" without referring to the previous status reported last along with the concrete interview date. An interview without preloads certainly has a higher chance of underreporting events. Sometimes, respondents even object the preload data and corrections of biographical spells collected in former waves are necessary, thus improving the quality of the self-reported biographical data.

The distinguishing feature of the NEPS is the combination of individual biographical data with individual attainments in competencies over the life course. Thus, it provides a data pool about individual competencies and their development over the life course on the one hand, as well as a data pool about educational processes and the unfolding of educational careers on the other hand. This constitutes a comprehensive and unique database to answer a great variety of research questions that are of high interest within the field of educational research.

5. Outlook

This article has discussed the benefits and challenges in longitudinal large-scale studies using SC4 of the NEPS as an example. Running a longitudinal large-scale study over several years is a highly demanding and sophisticated project. As shown before, such an endeavor requires a permanent awareness of the situation of the targets (that might even constantly change, especially when considering a certain age group such as the SC4, e.g. pathways from school to labor market or university) and a narrow monitoring of sample sizes in total and in subgroups. To this end, appropriate measures have to be used to cover diverse educational contexts (such as school tracks) and must constantly be refined, with a special focus on counteracting panel attrition and widening bias. Such demands involve the adaptation of instruments (e.g., including questions that are especially relevant to the targets or clearly connected to the study objective), the adjustment of incentives and communication strategies (e.g., targeted study information), and the modification of data collection procedures (e.g., shorter interviews or options for mode switch). Often measures that are suitable and well-tested for the purposes of cross-sectional studies are not in line with the longitudinal backbone: For longitudinal analyses regularly repeated questions are of high value as they facilitate very particular analyses

such as the study of life-course dynamics. However, from a participant's perspective answering the same questions again and again is tedious and often interpreted as a waste of time and energy. Here, questions focusing on uncomfortable contents are particularly critical. A plausible example is the repeated questioning of unemployed adolescents who fail to find a job and have to report this inability again and again, wave by wave. Thus, teams conducting longitudinal large-scale studies have to find a good balance between two crucial issues. First, the targets of a study have to be surveyed frequently enough to allow us to describe the particular life-course processes of interest accurately. Second, the gaps between survey waves should be wide enough to avoid panel attrition or nonresponse as a result of a lack of motivation.

The longer a study lasts the higher will be the value of the resulting longitudinal large-scale data sets. The obvious reason is the permanent accumulation of information concerning life-course developments. Moreover, in educational research there is a clear need for longitudinal large-scale assessments because competence development and educational trajectories can be monitored for longer time spans. Long-lasting studies are surely challenging, even if a lot of time and resources are invested in keeping a panel stable and in diminishing biasing effects. Hence, the design of longitudinal large-scale panels should not remain fixed over the whole course of the study but should be adapted from time to time. In this direction two main levers exist. The time spans between interviews can be widened substantially, possibly with panel care measures in between. Alternatively, a panel can be discontinued in order to invest the resources in a restart of a new panel with a fresh and unbiased sample. This is especially valuable when, for example, substantial changes have occurred in society or in settings that are of study interest, and which would otherwise be insufficiently covered.

6. References

Alderman, H., Behrman, J. R., Kohler, H. P., Maluccio, J. A., & Watkins, S. (2001). Attrition in longitudinal household survey data: Some tests for three developing-country samples. *Demographic Research*, *5*, 79-124.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data* (No. 46). Newbury Park, CA: Sage.

Allison, P. D. (2001). *Missing Data.* Thousand Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation models. *Journal of Abnormal Psychology, 112,* 545-557.

Allison, P. D. (2005, August). *Causal inference with panel data.* Paper presented at the Annual Meeting of the American Sociological Association, Philadelphia, USA. Retrieved from http://statisticalhorizons.com/wp-content/uploads/2012/01/Causal-Inference.pdf

Arjas, E., & Parner, J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, *31*, 171-187.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the life span within the German National Educational Panel Study (NEPS) – Editorial. *Journal for Educational Research Online, 5,* 5-14. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/viewFile/359/168

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., . . . Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), Education as a Lifelong Process - The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft*, *14*, 51-65. Wiesbaden: VS Verlag für Sozialwissenschaften.

Baltes, P. B. (1990). Entwicklungspsychologie der Lebensspanne. Theoretische Leitsätze [Life-Spam Developmental Psychology. Theoretical Principles]. *Psychologische Rundschau*, *41*, 1-24.

Baltes, P. B., Reese, H. W., & Lipsitt, L. P. (1980). Life-span developmental psychology. *Annual Review of Psychology*, *31*, 65-110.

Bela, D. (2016). Applied large-scale data editing. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study* (pp. 649-667) Wiesbaden: Springer VS.

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

Carstensen, C. H. (2009). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.). *Research on PISA. Research outcomes of the PISA Research Conference 2009 (*pp. 199-213). Heidelberg: Springer.

Coen, A. S., Patrick, D. C., & Shern, D. L. (1996). Minimizing attrition in longitudinal studies of special populations: An integrated management approach. *Evaluation and Program Planning*, *19*, 309-319.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). New York: Wiley.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer (Ed.) *Computerized adaptive testing: A primer.* (2nd ed., pp. 135–158). Hillsdale, NJ: Erlbaum.

Duchhardt, C., & Gerdes, A. (2013). *NEPS technical report for mathematics – scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 22). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXII.pdf

Elder, G. H. Jr., & Giele, J. Z. (2009). Life course studies: An evolving field. In G. H. Elder, Jr. & J. Z. Giele (Eds.), *The craft of life course research* (pp. 1-24). New York: The Guilford Press.

Elder, G. H. Jr., Johnson, M. K., & Crosnoe, R. (2003). The emergence and development of life course theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 3-19). New York: Kluwer Academic/Plenum Publishers.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2016, April). *Comparing linking methods – Rasch-scaled longitudinal competence data of the National Educational Panel Study (NEPS).* Paper presented at the 12th conference of the Austrian Psychological Association, Innsbruck, Austria.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/Survey%20Papers/SP_I.pdf

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*, 50-79. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/361/170

Haberkorn, K., & Pohl, S. (2013). *Cognitive basic skills (non verbal) – Data in the Scientific Use File* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/com_cogbasic2013_en.pdf

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XVI.pdf

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online, 5,* 162-188. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/365/172

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I. M., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online, 5,* 110-138. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/363/171

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics, 19,* 81-97.

Kalton, G. (2009a). Methods for oversampling rare subpopulations in social surveys. *Survey Methodology, 35,* 125-141.

Kalton, G. (2009b). Designs for surveys over time. In D. Pfeffermann, & C. R. Rao (Eds.), *Sample surveys: Design, methods and applications.* (Handbook of Statistics, Volume 29A) (pp. 89-108). Amsterdam: Elsevier.

KMK (2015). Definitionenkatalog zur Schulstatistik. Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. Kommission für Statistik [Definitions for school statistics: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. Comission for Statistics]. Retrieved from https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Defkat2015_2.pdf

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves (Ed.), *Survey nonresponse* (pp. 259-272). New Jersey: John Wiley & Sons.

Lynn, P. (Ed.). (2009). *Methodology of longitudinal surveys*. New York: Wiley.

Magnusson, D., & Bergmann, L. R. (Eds.) (1990). *Data quality in longitudinal research*. Cambridge: University Press.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*, 80-109. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/362/177

Nusser, L., Carstensen, C. H., & Artelt, C. (2015). Befragung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf Lernen: Ergebnisse zur Messinvarianz

[Questionnaires for students with special educational needs in the area of learning: Results from multi-group analysis]. *Empirische Sonderpädagogik, 7,* 99-116.

OECD (2014). *PISA 2012. Results: What students know and can do – Student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014), PISA, OECD Publishing.

Olczyk, M., Will, G., & Kristen, C. (2014). *Immigrants in the NEPS: Identifying generation status and group of origin* (NEPS Working Paper No. 41a). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXXXIa.pdf

Pfeffermann, D., & Rao, C. R. (Eds.) (2009). *Sample surveys: Design, methods and applications.* (Handbook of Statistics, Volume 29A). Amsterdam: Elsevier.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/366/173

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*, 1-25.

Ristau, I.-S., Meixner, S., Sixt, M., & von Maurice, J. (2015). Längsschnittliche Bildungsforschung in Deutschland – Herausforderungen und Umsetzung im Nationalen Bildungspanel am Beispiel der Startkohorte Klasse 9 [Longitudinal educational research in Germany: Challenges and implementations in the National Educational Panel Study considering the example of starting cohort Grade 9]. *Zeitschrift für Bildungsforschung*, *5*, 261-278.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. New Jersey: John Wiley & Sons.

Ruland, M., Drasch, K., Künster, R., Matthes, B., & Steinwede, A. (2016). Data-revision module – A beneficial tool to support autobiographical memory in life-course studies. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study* (pp. 367-384). Wiesbaden: Springer VS.

Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. New York: Springer.

Schöps, K., & Saß, S. (2013). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 23). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_XXIII.pdf

Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (2015). *The Education System in the Federal Republic of Germany 2013/2014. A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe*. Bonn, Germany: Standing Conference of the Ministers of Education and Cultural Affairs. Retrieved from https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-engl-pdfs/dossier_en_ebook.pdf

Senkbeil, M., & Ihme, J. M. (2012). *NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 17). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP-XVII.pdf

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online*, *5*, 139-161. Retrieved from http://www.j-e-r-o.com/index.php/jero/article/view/364/176

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: University Press.

Steinhauer, H. W., Aßmann, C., Zinn, S., Goßmann, S., & Rässler, S. (2015). Sampling and weighting cohort samples in institutional contexts: The National Educational Panel Study cohort samples of kindergarten children, students in grade 5 and in grade 9. *AStA Wirtschafts- und Sozialstatistisches Archiv, 9,* 131-157.

Steinhauer, H. W., & Zinn, S. (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 No. 2). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/Survey%20Papers/SP_II.pdf

Trahms, A., Matthes, B., & Ruland, M. (2016). Collecting life-course data in a panel design: Why and how we use proactive dependent interviewing. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study* (pp. 349-366), Wiesbaden: Springer VS.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: how to deal with missing data. *Journal of Clinical Epidemiology*, *55*, 329-337.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking competencies in horizontal, vertical, and longitudinal settings and measuring growth. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp.121-149). Göttingen: Hogrefe.

von Davier, M., & von Davier, A. A. (2010). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking (*pp. 225-242). New York: Springer.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), Education as a Lifelong Process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft*, *14*, 67-86. Wiesbaden: VS Verlag für Sozialwissenschaften.

Young, A. F., Powers, J. R., & Bell, S. L. (2006). Attrition in longitudinal studies: who do you lose? *Australian and New Zealand Journal of Public Health*, *30*, 353-361.

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). *The assessment of reading speed in grade 5 and grade 9. Status: 2012* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_rs_2012_en.pdf