**DIW BERLIN**

58

# Data Documentation

# BIOEDU (Beta Version): Biographical Data on Educational Participation and Transitions in the German Socio-Economic Panel Study (SOEP)

Henning Lohmann and Sven Witzke

**Data Documentation**  58

Henning Lohmann*

Sven Witzke **

**BIOEDU (beta version): Biographical data on educational participation and transitions in the German Socio-Economic Panel Study (SOEP)**

Berlin, September 2011

\*     Universität Bielefeld und DIW Berlin, Forschungsprofessor DIW Berlin. henning.lohmann@uni-bielefeld.de

\*\*    DIW Berlin, Forschungsbasierte Infrastruktureinrichtung SOEP. switzke@diw.de

# Contents

# Index of Figures

# 1        Introduction

The Socio-Economic Panel Study (SOEP) contains a broad range of variables including early childhood education and care, educational participation, educational degrees, and other related topics. However, the respective questions are contained in different questionnaires (e.g, individual questionnaire, household questionnaire, youth questionnaire) and the variables are not always in a format that is suitable for longitudinal analysis. For instance, transitions such as school enrollment or entry into tertiary education are not reflected in individual variables but can only be reconstructed by comparing the status in wave t with the status in wave t+1 (e.g., a transition to tertiary education took place if a person was not attending a university in wave t but was in wave t+1).[1] Generating such variables is time-consuming and prone to errors. It is the aim of the BIOEDU dataset to provide ready-made variables on educational transitions and related topics to facilitate longitudinal analysis. This documentation describes a beta version of the dataset (v27_0.2). If you have any comments or encounter any problems in using the dataset, please let us know.

# 2        Contents of the dataset BIOEDU

The BIOEDU dataset is based primarily on prospectively collected information. Therefore, it contains the most complete information on those individuals who were part of the survey population while they were attending school or other educational institutions. In total, the dataset contains information on 57,250 individuals, the portion of the SOEP sample that provided information on educational transitions and/or educational degrees. For the majority of this group, we have information on one educational degree only (n = 41,272). These are the persons who were not part of the sample when they were in school or higher education or when they experienced educational transitions. A smaller part of the sample is more interesting for the longitudinal analysis of educational participation. These are persons who

---

[1] But since 2010 a direct question on school enrollment is included in the newly introduced parental questionnaire for 7/8 years old children.

lived in a survey household while attending school, higher education, or training.[2]  Depending on the age of the individual, the dataset contains variables on:

- early childhood education and care (ECEC)
- entry into primary school
- transition to secondary school
- first exit from secondary school
- secondary school attendance after first exit from school
- first entry into and exit from vocational training
- vocational training after first exit from school
- first entry into and exit from tertiary education
- tertiary education after first exit from school
- highest educational degree attained and last observed educational participation

The variable *beinfo* summarizes which information is available for each person (see Section 4). We describe the construction of such variables in Section 3. In the next section, we provide some evidence on the validity of SOEP educational data by comparing them with official statistics. In Section 2.2 we discuss how SOEP data on education can be used in a longitudinal perspective.

## 2.1     Trends in educational attainment: Comparisons of SOEP data with other data sources

SOEP data provide a wide range of prospectively collected information, including indicators that are often not included in other large-scale datasets. They allow for longitudinal analysis at the level of both households and individuals. In this documentation, we focus on longitudinal analysis, in particular on analyses that take a person's context into account (see also Section 2.2). Nevertheless, we would like to know how the SOEP compares to other data sources, which, due to their sample size (if not register data), may provide more precise estimates of the distributions of educational variables in Germany.  The SOEP is a general household survey and therefore not directed at specific questions of educational research. Other data sources such as the German Mikrozensus or administrative registers might be better suited for providing an overview of the distribution of educational degrees or school

---

[2]      Accordingly, the first group is much older than the second group. At the time of the first observation in the sample, the first group was 45 years old on average while the second group was below 9 years on average.

attendance in the population in a cross-sectional perspective. We regard these comparisons as a test of the validity of the SOEP data. Therefore, before we describe the use of SOEP data to reconstruct educational biographies we give a brief overview of recent trends in educational attainment. We focus on participation in primary and secondary schools. Comparative results across different data sources may be added in later versions of this documentation.

Figure 1 shows the distribution of students according to the type of school in West Germany. The graph on the left side is based on administrative data ("Schüler-Klassen-Lehrer-Datensatz"); the graph on the right side uses SOEP data. The graphs differentiate among the following types of schools: 1. primary school, 2. lower secondary school (Hauptschule), 3. intermediate secondary school (Realschule), 4. upper secondary school (Gymnasium), 5. comprehensive school (Gesamtschule), 6. other types of schools. In addition, the administrative data allows for a separation of "schools with more than one track" (Schulen mit mehreren Bildungsgängen). These are schools that combine the lower two types of secondary school. As the SOEP questionnaire does not offer this option, it is most likely that these students are classified according to the chosen track within these schools (see also discussion in Section 3.3). Both graphs show similar trends over time. The share of students in primary school increased up to the mid-1990s and decreased afterwards (with a slight jump in the SOEP data at the end of the observation period). The share of students in upper secondary schools remained quite stable for a long time but has shown an increase in recent years. The share of students in the lowest type of secondary schools has been decreasing steadily. Since the end of the 1990s more students have been enrolled in intermediate schools while in earlier years the lowest type used to have the second largest number of students of all secondary schools. Comparing the beginning and end of the observation period, the share of students in intermediate schools has remained virtually unchanged while the share in comprehensive schools has increased steadily, albeit from a rather low level. Less than five percent of all students attend other types of schools. An even smaller share attends schools with more than one track.

Figure 1:
**Distribution of students by type of school, 1984-2007 (West Germany)**
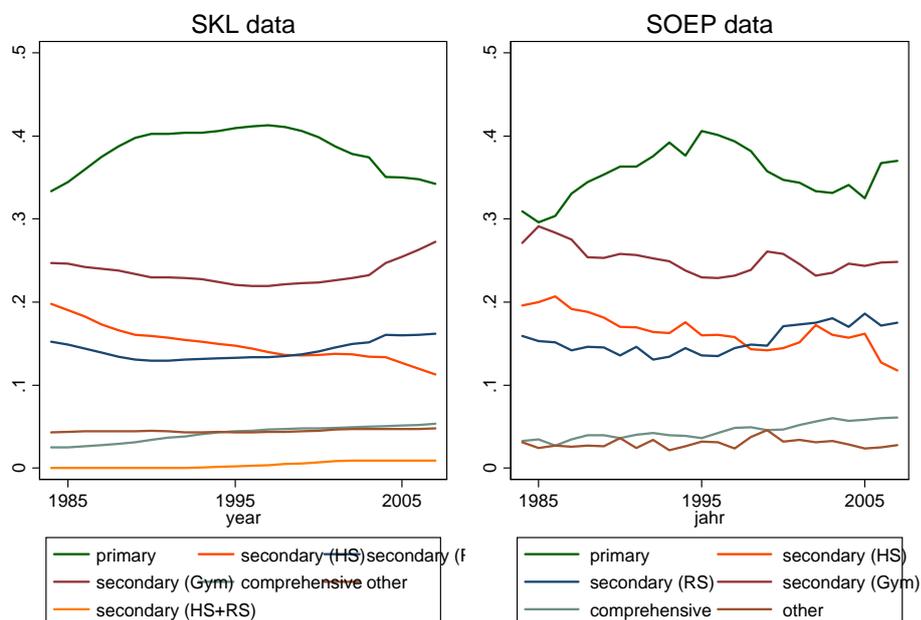


Notes: Types of secondary school: HS=Hauptschule (low), RS=Realschule (intermediate), Gym=Gymnasium (upper), HS+RS=Schule mit mehreren Bildungsgängen (combined low and intermediate), Sources: SOEP 1984-2007 (weighted), Schüler-Klassen-Lehrer-Datensatz (KMK 1994, div. years).

East Germany deviates from this pattern—mostly due to the higher prevalence of schools with more than one track (Figure 2). Looking at the results based on the administrative records, we can see that about 15 to more than 20 percent of all students in East Germany attend this type of school, that only a negligible percentage attend lower track secondary schools, and that the share of students in intermediate schools has been in steep decline since 1997. On the basis of the SOEP data, we can only compare the share of students in upper secondary schools to the two lower school types, as no information on schools with more than one track is available. When combining the shares for the two lower types of secondary schools, we find similar developments as reported for the three tracks individually (including schools with more than one track) in the administrative data. However, the decline in the share of students in the lower and intermediate schools is larger, while the increase in students in the upper secondary school has not flattened out in recent years. In both graphs, we observe a steep decline of the share of students in primary school, which reflects the marked changes in fertility in East Germany since reunification.

Figure 2:
**Distribution of students by type of school, 1992-2007 (East Germany)**



Notes: Types of secondary school: HS=Hauptschule (low), RS=Realschule (intermediate), Gym=Gymnasium (upper), HS+RS=Schule mit mehreren Bildungsgängen (combined low and intermediate), Sources: SOEP 1984-2007 (weighted), Schüler-Klassen-Lehrer-Datensatz (KMK 1994, div. years).

Figure 3 provides evidence on the development in the share of students according to type of school in Germany (East and West) since 1992. Given the much larger share of the population living in West Germany, the result is similar to that presented in Figure 1. However, there is also a decline in the share of students in primary school, which is driven by developments in East Germany. By and large, SOEP and administrative data show, at least at this level of aggregation, fairly similar trends. As seen above, in the case of East Germany, we have to take into account the higher prevalence of schools with more than one track, which is not appropriately reflected in the SOEP data. One solution to this problem would be to simply add up the shares of the lower and intermediate secondary schools. Apart from the issue of a higher prevalence of schools with more than one track schools in East Germany, the SOEP-based graph presents a similar picture to the administrative data.
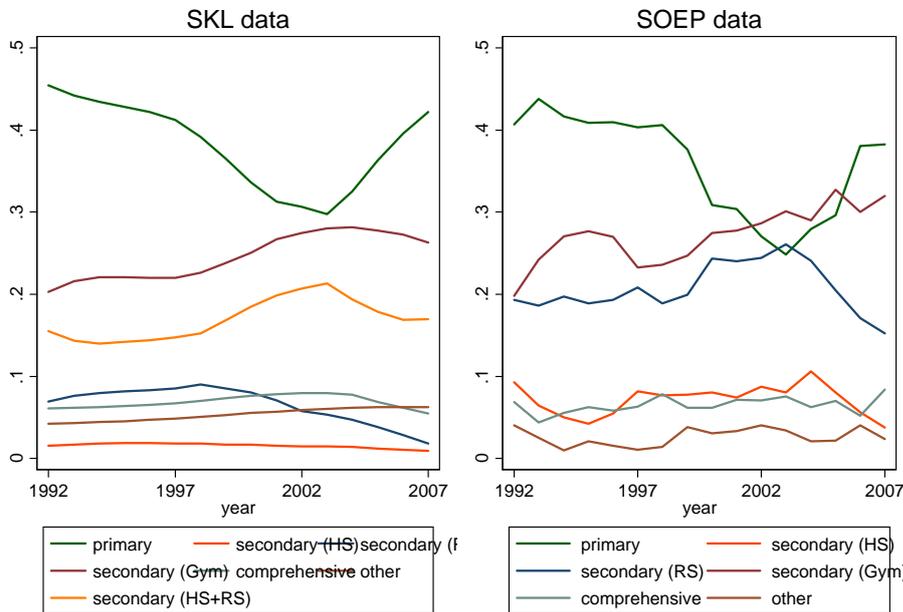
Figure 3:
**Distribution of students by type of school, 1992-2007 (Germany)**
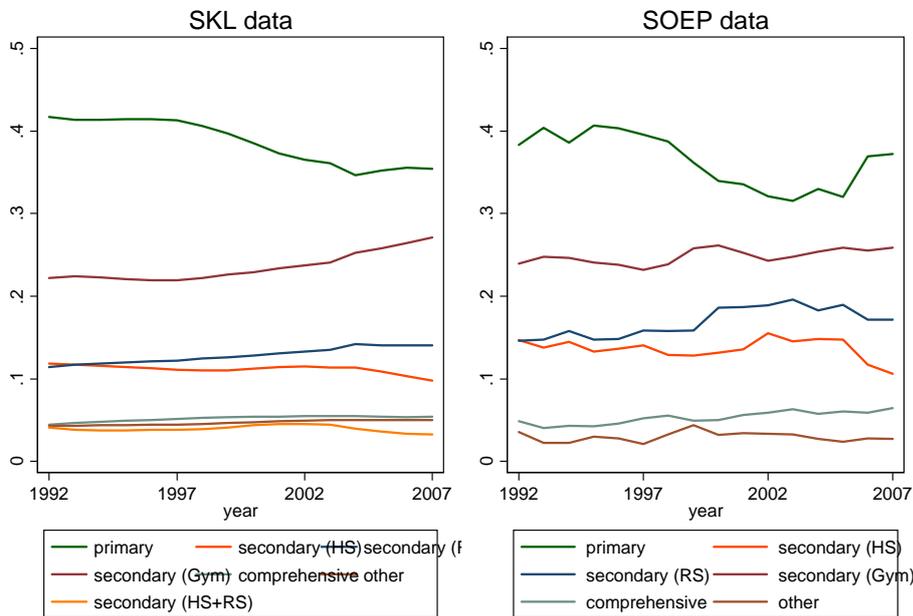


Notes: Types of secondary school: HS=Hauptschule (low), RS=Realschule (intermediate), Gym=Gymnasium (upper), HS+RS=Schule mit mehreren Bildungsgängen (combined low and intermediate), Sources: SOEP 1984-2007 (weighted), Schüler-Klassen-Lehrer-Datensatz (KMK 1994, div. years).

## 2.2      Life course perspective

In Section 2.1 we compared SOEP data with administrative data in a cross-sectional perspective. However, the main objective of generating the BIOEDU dataset is to facilitate longitudinal analysis. The basic idea is to provide information on the educational life course starting from the entry into early childhood education and care through primary, secondary, and tertiary education and training in later life. Given the duration of more than 25 survey years, the SOEP data cover rather long stretches of the educational life course for a number of individuals. This perspective is illustrated in Figure 4.

Figure 4:
**Educational life course**
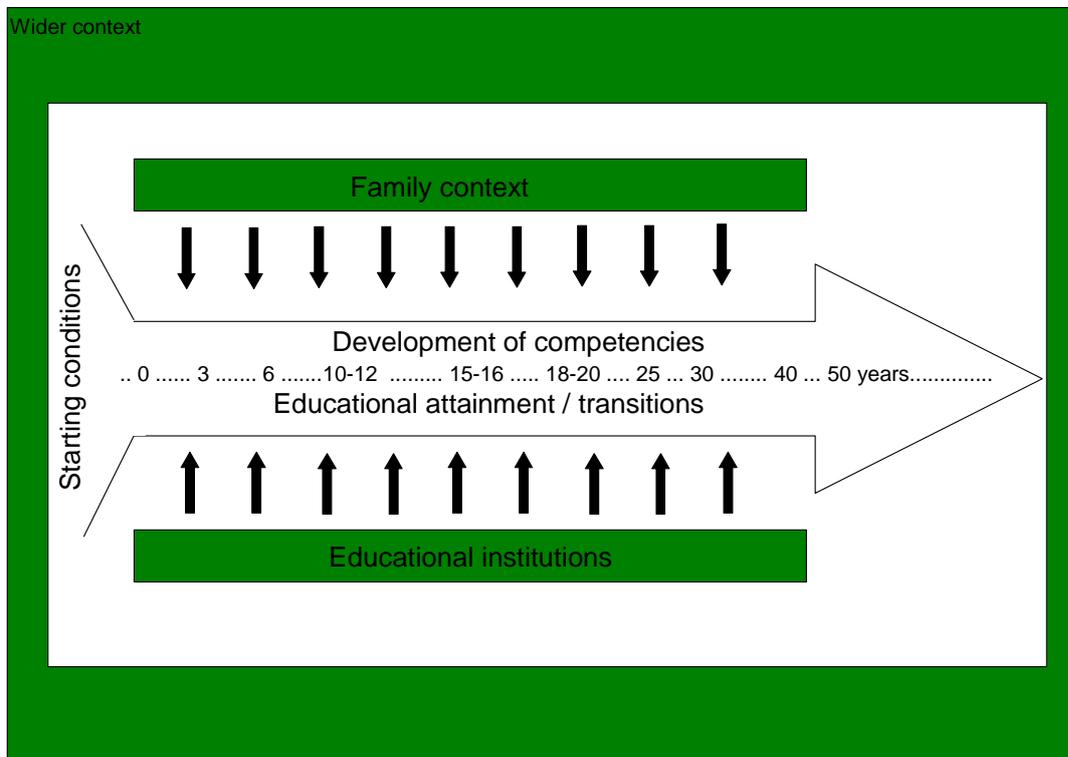


Source: Lohmann et al. 2009: 262.

The BIOEDU dataset provides information on educational participation and educational transitions. Since it is based on prospectively collected information, the duration of the observed portion of the educational life course differs individually. Yet individuals are not observed in isolation, but within the context of their households and families, which are also observed prospectively. In addition, the SOEP provides some information on the context of educational institutions (such as private or public, distance to closest secondary school). However, compared to the information on the household context, such data are rather limited. What we would like to stress is that the BIOEDU dataset provides variables on the educational life course, which can, via the usual identifiers (e.g., persnr), be matched to all other SOEP data, which allow for the analysis of the educational life course within the context of households and families (and wider contexts). In particular, we focus on crucial transitions such as the start of secondary school or post-secondary education. The SOEP has often been used to analyse such transitions. However, the reconstruction of transitions is a tedious task, which also requires assumptions on how to deal with inconsistencies. Furthermore, it is often very difficult to document the individual steps that have been taken to achieve consistency in the data. Therefore, probably every researcher who uses SOEP data for the analysis

of educational transitions uses a somewhat different concept of transitions. The aim in generating the BIOEDU dataset is to provide a commonly used and fully documented basis for the analysis of transitions over the full duration of the educational life course.

## 2.3 Panel data and spell data

The SOEP contains different types of information that we use to reconstruct the educational life course. These are:

1. *Data on current educational participation*: Respondents provide information on their own current educational participation (or parents provide information on their children) in the yearly individual, youth, or household questionnaire. These data cover early childcare and education, primary and secondary school, vocational training, and tertiary education, and are used for the reconstruction of all types of educational transitions.

2. *Data on degrees obtained since the beginning of the last survey year or dropouts (since 2000)*: Adult respondents are asked every year whether they completed any type of education or obtained a degree in the last year. These data are used in addition to the data on educational participation to reconstruct the exit from secondary school, the exit from vocational training, and the exit from tertiary education.

3. *Further data on educational degrees*: Information on educational degrees is also collected in the respondent's first interview and the biographical questionnaire. These data usually refer to degrees obtained before a respondent joined the panel population. In addition, data on educational degrees were collected from all respondents in 2000.

4. *Data from the activity calendar*: Every year, adult respondents complete the "activity calendar," which covers every month of the year prior to the survey. Currently it differentiates among 12 activities such as full-time or part-time work but also school or vocational training. These data are available as monthly variables and in spell format (cumulated over all years) in the dataset ARTKALEN. Using the education spells provided in ARTKALEN would be an alternative to reconstructing the educational life course. Still, we mainly rely on the yearly panel data for two reasons: 1. Respondents complete the activity calendar for the first time in the year when they turn 17 (18 since 2006). Early childhood education and care, primary school, the transition to secondary school, and—for those who left school at 15 or 16—even the exit from

school and the entry into vocational training are not covered in the activity calendar.
2. The categories of activities are too broad. Up to 1999, in-house (firm-based) vocational training and further training fell into one category. Still, general schools, vocational schools, and universities fall into one category. In many cases, it is not possible to separate when school ends and university starts. Nevertheless, we use the activity calendar as an additional data source for the reconstruction of some of the transitions (see Sections 3.7, 3.9 and 3.10).[3]

5. *Additional data*: We use data on job changes and on the current occupation as additional criteria when reconstructing entries into vocational training (see Sections 3.6 and 3.7)

The "mother-child" and "parent-child questionnaires" (see the respective chapters in Frick/Lohmann 2010) provide additional information on childcare and educational participation. However, the respective questionnaires were first introduced in 2003 (for children born in 2002 and later). In addition, they do not provide an annual measure for each child. Therefore, we decided not to use these data. However, for researchers interested in childcare and child development, these data provide a valuable source of information on many issues not covered in the BIOEDU dataset.

# 3 The reconstruction of crucial transitions in the educational life course: The construction of variables

The SOEP, as a general household panel study, is not specifically directed at the analysis of educational life courses. Nevertheless, since the beginning of the panel in 1984, the survey has contained questions on the educational attainment of all respondents (aged 17 and older) and of children younger than 17 years living in survey households. After more than 25 years of survey duration, these data provide an extremely valuable source for the reconstruction of educational life courses. In the following, we describe how we use these data to reconstruct educational transitions starting before school enrollment and extending up to post-secondary education.

---

[3] In addition, the dataset PBIOSPE contains information on all activities from the age of 15 to entry into the panel population. However, the episodes are recorded on a yearly basis and the categories are even broader than in the ARTKALEN data, which makes it almost impossible to reconstruct transitions from general school to university and other types of education.

The reconstruction of these transitions is primarily based on yearly information on educational participation (i.e., entry and exit reconstructed from changes in participation). For later transitions, further information is derived from explicit questions in the questionnaires about the completion of secondary schooling, vocational training, and tertiary education (changes in the year prior to the survey for persons aged 17+ years only; exception in the youth questionnaire: completion of schooling before age 17). In the following, we describe in detail how each of the variables is constructed.

One remark on the variable naming conventions: the variable names always begin with "be" which stands for "biography of education" (in analogy to other biographical datasets). The third and fourth letters denote the type of transition or similar. For instance, t0 stands for variables on the first and t1 for the last year in child care. Variables on starting school contain a t2 and so on (up to t8 = exit from tertiary education). Variables containing an x as the third letter contain information on the last observed year in education or on the highest educational degree obtained (x4, x6, x8).

## 3.1    Early childhood education and care (ECEC) [bet0…., bet1….]

Data on enrollment of children living in a survey household in ECEC are obtained from the household questionnaire.[4]  Currently, in the dataset BIOEDU, we do not distinguish among different forms of ECEC services, which include nursery school, kindergarten, and daycare centers (including after school care). We do also not distinguish between part-time and full-time care. The reason is that the respective questions have changed over time and it is therefore difficult to provide consistent information. Thus, if the variables described below indicate that a child has been in ECEC in a given year, this means that he or she has been in one of the aforementioned institutions for at least a few hours per day.[5] In contrast to later transitions such as school enrollment or the transition to secondary school, entry into and exit from primary education is less institutionalized. Therefore, we do not try to reconstruct variables on entry and exit. Instead, we provide information on the first and last observed year of attendance of ECEC services. A disadvantage of this approach is that a given year may not necessarily be the first or last year of attendance. This is only the case if all years of at-

---

[4] For children born 2002 and after, additional and more detailed information is collection in the mother-child and parent-child questionnaires (see Section 2.3).

[5] Before 1995, we cannot separate institutional childcare from care provided by childminders (outside the parental household). Up to 1994, the variable indicates that a child has been in either of the two types of care.

tendance have been observed. In many cases, the observation window does not start at birth but at a higher age. We cannot be sure that the first *observed* year is in fact the first year of ECEC (because earlier years may not be observed). Therefore, the variables should not be used without taking into account the issue of left- or right-censoring. To facilitate the handling of this issue, the dataset BIOEDU contains two variables (firstobs, lastobs) which record the age when a person is observed for the first or last time in the panel. In addition, the variables bet0obs and bet1obs refer to non-censored entries only.

The dataset contains the following variables on ECEC:
- bet0obs: first year of ECEC observed (uncensored)
- bet0year: first observed year of ECEC
- bet0age: age (in years), first observed year of ECEC
- bet0agem: age (in months), first observed year of ECEC
- bet0fst: state, first observed year of ECEC
- bet1obs: last year of ECEC observed (uncensored)
- t1year: last observed year of ECEC
- t1age: age  (in years), last observed year of ECEC
- t1agemo: age  (in months), last observed year of ECEC
- t1fst: state, last observed year of ECEC

## 3.2     Entry into primary school [bet2….]

For most of the sample, explicit questions on school enrollment have not been asked.[6] However, every household reports the current educational participation of all children younger than 17 years. This information is used to construct variables on school enrollment. Basically, we assume that school enrollment has taken place if a child is in school in a given year and was not in school in the previous year. But this will yield reliable information on school enrollment only if educational participation is fully recorded and if a child has not temporarily dropped out of school. While the latter is a rare case, missing or implausible information is more often a problem. Therefore the construction of the enrollment variable is based on a basic algorithm followed by a series of checks, which are described in the following.

---

[6] A similar question was introduced in 2010 as part of the parent-child questionnaire for 7/8-year-olds (see Section 2.3).

*basic algorithm:*

If a child attends primary school[7] in a given year and was not in school the year before, we assume that school enrollment took place in the intervening period. Since school usually starts in late summer, we assume that school enrollment took place in the last August prior to the interview (e.g., if the first interview took place in March 2007 and the second in April 2008, we assume that school enrollment took place in August 2007). Using this basic algorithm, we reconstruct 5,967 cases with observed school enrollment.

*additional checks:*

1. *Deletion of double entries*: For 91 children, we reconstructed two entries into school. In most cases this was due to missing data in previous years. Therefore, we used a first criterion to delete implausible entries into school. If the assumed year of school enrollment was not the first year of school, it was not regarded as the year of enrollment. For instance, if primary school attendance was observed in 2005 and in 2007 but not in 2006, we would, using the basic algorithm, generate two entries into school (in August 2004 and 2006). To correct for virtual double entries, new observations of school attendance (i.e., in school in a given year but not the previous year) are not counted as school enrollment if school was attended in previous years. This reduces the number of observed school enrollments to 5,932 and leaves no cases with two observed enrollments.

2. *Application of age thresholds*: A second criterion is the age at school enrollment. There are age thresholds that define the earliest and latest entry into school. However, the respective laws differ by state and over time. Furthermore, depending on the child's development, there is scope for discretion in most states. We therefore use a rather wide age margin to exclude only those cases that are most likely to be flawed.[8] We delete reconstructed entries if school entry took place in the year a child turned four years old and if the child did not attend school in the following year (or if an assumed entry took place at an even younger age). The rule for the upper age limit is as follows: We delete reconstructed school enrollment if a child turned eight years old in the year of school enrollment and if he or she attended secondary school

---

[7] In a few cases, school enrollment took place in some other type of school, such as Waldorf schools, comprehensive schools, which sometimes start at grade 1, special education schools, or the GDR type of comprehensive school starting at grade 1 ("Polytechnische Oberschule", applies to the year 1990 only). The type of school at the time of enrollment is recorded in the variable bet2type.

[8] Age is computed based on the year when a child turned four, etc.: we cannot compute the exact age of every child as the month of birth is missing in a number of cases.

one or two years after school enrollment. If a child was even older (nine years or above) at the time of school enrollment, we regard this information as unreliable and delete the entry without further assumptions. Using the age margins reduces the number of observed school enrollments to 5,849.

3. *Restoring erroneously deleted entries*: In some cases, the application of the first or second criterion results in a deletion of two observed school enrollments (leaving none). Take, for instance, a case where the parents erroneously report attendance of primary school (instead of some type of pre-school) at age three and correctly at age six. Based on the basic algorithm, we observe two entries to school. Based on the first criterion, we would delete the second one (age six) because there is information on school attendance in an earlier year. But we would also delete the first observation of school enrollment because it is outside the age margin (second criterion). Therefore we check whether there was a second observed school enrollment at a more plausible age if we delete observations based on the second criterion. This is not often the case. Instead of 5,849, we observed 5,875 school enrollments.

The reconstruction of school enrollment is documented in a variable that contains information on all observed irregularities (bet1info; see Appendix II). Bet2info captures whether a case has been corrected as described above. However, in about 98 percent of all cases, the school enrollment variable has been created without any further corrections. We regard these cases as the most reliable, but also consider the information on the majority of other cases to be highly reliable. If you disagree, you could exclude such cases or use the information captured in the variable bet1info to check them on your own. In addition, the variable captures whether there are any other factors that may affect the reliability of the measure (month of interview around August, birth month of child missing).[9] This includes cases missing a full school year due to the timing of the yearly interviews. Usually the interviews are carried out in spring but there are some exceptions in which the interview did not take place until August. Take as an example a household that was interviewed in April 2006, in September 2007, and in April 2008. In this case, we have information on the 2005/06 school year (first interview) and on the 2007/08 school year (second and third interview). Information on the 2006/07 school year is missing. Such "irregular" patterns are also recorded in the variable bet2info.

---

[9] If the interview took place in August, we cannot be sure whether the information on school attendance referred to the school year that just started or the previous one. However, since most of the SOEP interviews take place in spring we do not face this problem often. If the month of birth of the child is missing, the age at school enrollment can only be computed approximately, which lowers the reliability of checks based on the age of children.

The dataset contains the following variables on school enrollment:
- bet2obs: school enrollment observed
- bet2year: year of school enrollment
- bet2age: age (in years) at school enrollment
- bet2agem: age (in months) at school enrollment
- bet2type: type of school at school enrollment
- bet2fst: state, school enrollment
- bet2info: documentation of reconstruction of school enrollment

## 3.3 Transition to secondary school [bet3…]

As in the case of school enrollment, we reconstruct transitions to secondary school using the variables on children's educational participation (household questionnaire). Basically, we assume that a transition to secondary school has taken place if a child is in secondary school in a given year and was in primary school in the previous year. This approach yields unambiguous results only if school attendance is fully recorded and we are always able to differentiate between primary and secondary schools. As this is not always the case, we apply additional checks to reconstruct the transitions. Before we explain these checks in more detail, we briefly describe two main reasons why it is sometimes difficult to distinguish between primary and secondary schools: First, there are schools such as Waldorf or special education schools that contain both primary and secondary schools. In most survey years, these schools are recorded as "other schools," which makes it impossible to say whether the school is a primary or a secondary school. Second, the German educational system, although still highly stratified, still includes comprehensive schools and other schools with similar aims. However, there are a wide variety of such comprehensive elements in the German educational system, mostly due to the fact that education is the responsibility of the states. Thus, there is not one German educational system but sixteen. In addition, changes have taken place over time. As it is impossible to capture the details of each of these systems in a country-wide questionnaire of a general household survey, there is sometimes a certain ambiguity in how specific types of schools or programs are classified.

In most survey years, the SOEP questionnaire on general school attendance offers six categories: primary school, lower secondary school (Hauptschule), intermediate secondary school (Realschule, Mittelschule), upper secondary school (Gymnasium), comprehensive

school (Gesamtschule), and other schools (in some years there is an additional category for special education schools). This outline largely corresponds to the traditional (West-) German educational system, with its four years of primary school, three types of secondary schools, and comprehensive and other schools as additional non-standard options. However, there are several elements of destratification in grades five and six. In some states (primarily Berlin and Brandenburg), primary schools cover the first six grades.[10] In many Western states, there are elements of comprehensive school in grade five and six usually known as the "orientation stage" or "transition stage" (Orientierungsstufe, Förderstufe, Erprobungsstufe). Given the aim of prolonging joint learning, the "orientation" or "transition stages" can be regarded as a part of primary school. But in many states, these "stages" are integrated into the schools in the three-track system. In other states, they are institutionalized as a separate type of school (in a certain sense similar to comprehensive schools). The most relevant case is the state of Lower Saxony, where these separate schools existed from the 1970s up to 2004. Therefore, there is some ambiguity in which type of school parents might choose when they answer the questions on school attendance if their children are enrolled in the "orientation" or "transition stage." Today, in most states, the system has been abolished but as the SOEP data date back into the mid-1980s these institutional settings have still to be taken into account in the reconstruction of educational transitions into secondary school.[11] In addition, there is a trend towards schools that combine lower and intermediate secondary schools (see also Section 2.1). In some states, these are the only type of schools apart from the Gymnasium (upper secondary). In other states, these schools exist alongside the traditional three-track system. Unfortunately, there are many different names for these types of schools (Mittelschule, Oberschule, Sekundarschule, Regelschule, erweiterte Realschule, Stadtteilschule, general term: "Schule mit mehreren Bildungsgängen" – schools with more than one track). Again, some ambiguity in respondents' answers must be expected since some might regard these schools as comprehensive schools, others as lower or intermediate secondary school or simply as other types of schools. Therefore, the data is expected to be most reliable in states where the standard is still the three-track system. In other states, it is advisable not to focus primarily on the differences between the two lower types of secon-

---

[10] In Berlin, as an additional exception, some of the secondary schools of the highest track start at the fifth grade ("Grundständiges Gymnasium") as in most other federal states.

[11] Bavaria represents another specific case. Until 1999, the lowest and medium tracks were merged during grade 5 and 6 ("Teilhauptschule") and students were separated in grade 7. Only the highest track started in separate schools at grade 5. There are some other minor or not as long-lasting exceptions, such as the introduction of an orientation phase in Mecklenburg-Western Pomerania starting in the academic year 2006/07.

dary school but on the differences between these two ("Hauptschule", "Realschule" or "Mittelschule") and the upper secondary school ("Gymnasium").

*Basic algorithm:*

The basic approach is that a transition to a secondary school has taken place if a child is in secondary school in a given year and was in primary school in the previous year. Using this simple algorithm, we observe 6,150 cases with transitions to secondary school. In additional steps, we check for double or implausible transitions.

*Additional checks:*

1. *Deletion of double transitions*: Using the basic algorithm, we reconstructed multiple transitions for a number of cases (396 cases with 2 transitions, 11 cases with 3 transitions). Such multiple transitions occur, for instance, when parents report that their child attended primary school in the first year, secondary school in the second, primary in the third and secondary in the fourth. We assume that if a transition to secondary school has taken place, there is no transition back to primary school (probably ignoring some rare exceptions).[12] Hence, we deleted reconstructed transitions to secondary school if a child attended secondary school before the assumed transition. Applying this rule, we significantly reduced the number of cases with multiple transitions (to just eight). This does not, however, affect the number of cases for which we observe a transition (n=6,150).

2. *Attendance of primary school after transition*: In a second step, we deleted all transitions if primary school was attended after the assumed transition. This reduced the number of cases with an observed transition to 6,122.

3. *Attendance of primary school too short*: Given the compulsory four or six years of primary school (depending on the state), we considered transitions occurring two or three years after school enrollment to be unreliable. Applying this criterion (where the age of school enrollment was observed), we reduced the number of observed transitions to 5,957.

4. *Attendance of primary school too long*: We also deleted reconstructed transitions that took place three or more years after the usual point of transition. For instance, a

---

[12] It is likely that such transitions to primary school occur when a family moves to Berlin or Brandenburg, the two states in which primary school lasts up to grade six (all other states: grade four). Hence, a child which attends grade five (secondary school) in any other federal state would be obliged to move back into primary school in Brandenburg or Berlin. However, a thorough check of the data reveals that residential mobility does not explain the apparent fact that a substantial number of children attend primary school after having already attended secondary school.

child is enrolled at age six in states with four years of primary school. Without grade repetition, we would expect the child to enter secondary school at age 10. If, in this example, the child were 13 years or older, we would regard the transition as unreliable. Applying this rule (where the age of school enrollment was observed), we deleted 20 transitions as unreliable, leaving 5,937 transitions.

5. *Application of age thresholds*: We also deleted transitions that occurred at an age well below or above the expected age at transition. This reduced the number of cases with observed transitions to 5,830.

6. *Residual cases*: Applying the rules described above to the data, we still observed two assumed transitions for one case. We deleted the second of these transitions.

7. *Restoring erroneously deleted entries*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible transitions. An example would be a child that, according to the parent, attended primary school at age 8, secondary school at age 9, primary school at age 10, and secondary school at age 11. Using the basic algorithm, we would reconstruct two transitions: one between age 8 and 9, and a second between age 10 and 11. Following our checks, we would delete the latter transition because the child appears to have attended a secondary school before this transition (Step 1, see above). But we would also delete the former transition because it must have taken place between age 8 and 9, which is implausibly early. Now the reason for deleting the transition between age 10 and 11 is no longer relevant. We therefore restored the information on the later transition. After this last step, we have reconstructed the transition to secondary school for 5,851 cases.

The fact that a transition which has been generated using the basic algorithm has been deleted or restored is documented in a variable (bet3info, see Appendix II). In 88 percent of the cases the transition has been reconstructed using the basic algorithm only, i.e., without further corrections based on additional assumptions. In the majority of the cases where corrections have been made, they have been made due to inconsistent information. The assumptions explained above allow us to increase the consistency of the data. We are fully aware that such an approach is prone to errors. But our objective is to provide data that seem plausible after the thorough checks we have applied. If you are not confident in all of the assumptions that we use to run the checks, you can identify the cases affected and exclude them or use the original data to reconstruct the transitions on your own.

In the construction of the variables that capture transitions to secondary school, we cannot, as explained above, reflect the full complexity of the German educational system or the changes that have occurred in it over time. However, we take into account some very crucial differences between the states and changes in them over time. First, in Berlin and Branden-burg, there are six years of primary school (and other, less prevalent forms of prolonged joint learning in some states). In these states, other rules concerning age at transition and duration since school enrollment apply. Second, in Lower Saxony up to 2004, the orientation stage existed as an additional type of school. We assume that the transition to secondary school took place after the orientation stage. Third, in Bavaria, a combined lower and inter-mediate secondary school existed in grades five and six ("Teilhauptschule"). After a change in the respective state legislation, this model was gradually phased out starting in 1999.

The dataset contains the following variables on transitions to secondary school:
- bet3obs: transition to secondary school observed
- bet3year: year of transition to secondary school
- bet3age: age (in years) at transition to secondary school
- bet3agem: age (in months) at transition to secondary school
- bet3type: type of school, first year of secondary school
- bet3fst: state, transition to secondary school
- bet3info: documentation of construction of variable

## 3.4      First exit from secondary school [bet4….]

The minimum duration of compulsory schooling is nine years in Germany, and in some states longer. Assuming school enrollment at the age of six, students can exit secondary school on the shortest track at age 15. Students graduate from the highest track after 12 or 13 years of schooling, i.e., at the age of 18 or 19 (or later in the case of grade repetition or delayed school enrollment). The fact that typical school exit ages range from 15 to 19 years is crucial for the construction of a school exit variable with the SOEP data. In the case of younger school leavers, school exit takes place before the person becomes a respondent in the sur-vey. Others are already respondents at the time they leave school. As a consequence, in the former case, it is only possible to derive prospectively collected information on school leav-ing from the question on children's school attendance in the household questionnaire. Simi-lar information exists for adult respondents. Therefore, in a first step, we reconstruct the

year of school exit using information on school attendance only.[13] In a second step, we use information on school leaving certificates, which is only available for adult respondents.

*basic algorithm (Step 1):*

On the basis of yearly information on educational participation, we construct a variable of school exit. We compare school attendance in t and t+1. If a person attends a secondary school in t and no school or a post-secondary school in t+1, we assume that an exit from secondary school took place in the intervening period.[14] We consider only the first exits from school (but see Section 3.5). Using this simple algorithm, we observe an exit from school for 6,113 cases.

*additional checks (Step 1):*

1. *Deletion of double exits*: We deleted exits from school if they appeared not to be the first exits (e.g., vocational training or tertiary education before assumed exit from school). This left 6,024 cases with an observed first exit from school.

2. *Inconsistencies in history of school attendance*: Gaps due to missing data sometimes make it difficult to determine the exact year of leaving school. We therefore checked cases with missing data thoroughly. If we found inconsistencies in the history of school attendance (i.e., not at least four years of secondary school attendance observed), we deleted the assumed transition as unreliable. This left 5,924 cases with an observed exit from school.

3. *School attendance implausibly long or short*: We also deleted an assumed transition if the observed duration of secondary school attendance was implausibly short or long (measured as duration since transition to secondary school). This left 5,911 cases with an observed exit from school.

4. *Application of age thresholds*: If a person was younger than 12/13 or older than 23/25 at the time of the assumed exit from school, we regarded the exit information as unreliable. After the deletion of such unreliable transitions, we observed an exit from secondary school for 5,822 cases.

---

[13] For the reconstruction of exits from secondary school we do not use the activity spell data contained in the datasets ARTKALEN or PBIOSPE (see Section 2.3). Although the datasets also contain valuable information on educational participation, there are a number of caveats why we do not use these data: PBIOSPE: advantages: starts at age 15, disadvantages: yearly data, very broad categories ("school, tertiary education and evening school"), since 2000, no longer answered by first-time respondents at age 17; ARTKALEN: advantages: monthly information, disadvantages: starts at age 16, broad categories ("school and tertiary education"). In general, the use of SOEP activity spell data would require consistency checks with the annually collected data. As the data does not promise much additional information for the reconstruction of first exits from school, we refrained from going in this direction.

[14] In a few cases we used information from the "gap"-datasets ($PLUECKE).

5. *Residual cases*: Applying the rules described above to the data, we still observed two assumed first exits from school for a number of cases. In all such cases, we deleted the second of these transitions.

6. *Restoration of erroneously deleted exits*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible transitions. An example would be a person who, according to the data, had attended secondary school at ages 13 and 15 and a vocational school at ages 14 and 16. Using the basic algorithm, we would reconstruct two transitions: one between age 13 and 14 and a second between age 15 and 16. Following our checks, we would delete the latter transition because the person apparently attended a vocational school before this transition (see above). But we would also delete the former transition because of the age restriction we apply. In this case, the reason for deleting the second assumed transition is no longer relevant. We therefore restored the information on the respective transition. After this last step, we observed the first exit from school for 5,870 cases.

The fact that a first exit from school generated using the basic algorithm was deleted or restored is documented in a variable (bet4info, see Appendix II). In more than 95 percent of the cases, the transition has been reconstructed using the basic algorithm only, i.e., without further corrections based on additional assumptions.

*basic algorithm (Step 2):*
In a second step, we add information on school leaving certificates. Such information is collected in various ways:[15]

- Prospectively collected data on completion of education (year prior to the survey): Every year, all respondents aged 17 and over are asked whether they have completed education since the beginning of the previous survey year and what type of school leaving certificate they obtained (since 2000: if they dropped out without graduating).
- Retrospective data on educational degrees: Usually in the first or second year of survey participation, new respondents complete the biographical questionnaire,

---

[15] Apart from the following options, up to the early 1990s there existed specific questionnaires for foreigners and East Germans. The respective questions were included in the general questionnaires in subsequent years.

which covers topics such as the respondent's social origins and life course up to entry into the SOEP. This includes questions on educational degrees.

- "Inventory" in 2000: The variables on the educational degrees are updated on a yearly basis if new educational degrees have been obtained (in the year prior to the survey). In addition, in the year 2000, every respondent in the sample was asked again about his or her educational degree. Unfortunately, educational degrees are not always consistent with the information collected in earlier interviews.

- New respondents: Since 1998, every first-time respondent has been asked about his or her educational degree, as the information which was collected via the biographical questionnaire is usually not available in the first wave.

Our basic approach is to match the information on school leaving certificates to the first year of exit from school, which is reconstructed using information on school attendance (see Step 1). As described above, for many respondents, we have information on more than one school leaving certificate. This could refer to the same certificate or different ones if a person attended school between the two points of observation. In order to use the information on the certificate that matches the first exit from school, we use information collected at the time of the exit from school, which is, however, not always available. As a basic rule, we prioritize variables that contain precise information on the exact time when a certificate was obtained:

1. prospective data on the school leaving certificate in year of first exit from school
2. retrospective data on last school leaving certificate in year of first exit from school
3. prospective data on dropping out of school (information available since 2000) in year of first exit from school
4. prospective data on school leaving certificate in year after first exit from school
5. prospective data on school leaving certificate in year before first exit from school
6. retrospective data on last school leaving certificate in year before or after first exit from school
7. retrospective data on last school leaving certificate collected the year after first exit from school if no school attendance after first exit from school was observed

We provide information on the source of the information on the school leaving certificate and the year when it was collected (variables bet4cert_s and bet4cert_g).[16] In almost three quarters of all cases, the variable is based on the prospectively collected information (personal questionnaire). In about 85 percent of the cases, the information was collected in the year of the exit from school or in the subsequent year. If no matching data is available, the type of school in the last year of participation (bet4type) can be used as a proxy for the school leaving certificate (bet4cert). The fact that no school leaving certificate is recorded does not necessarily mean that a person has dropped out of school. Our sample contains more than 10 percent of school leavers without information on a school leaving certificate. This figure is higher than the share of school drop-outs in Germany. In particular, it is often not possible to distinguish school leavers from school drop-outs if the event occurred at age 15 or 16, i.e., at an age when a person is not a respondent. In particular, for survey years before 2000, it is difficult or even impossible to clearly distinguish school leavers from school drop-outs. The situation improved after a question on school drop-out was included in the questionnaire and with the introduction of a youth questionnaire for 17-year-olds.

The dataset contains the following variables on the first exit from school:
- bet4obs: first exit from secondary school observed
- bet4year: year of first exit from secondary school
- bet4age: age (in years) at exit from secondary school
- bet4type: type of school, last year of secondary school
- bet4cert: school leaving certificate
- bet4cert_s: school leaving certificate (data source)
- bet4cert_g: school leaving certificate (information gathered … years after leaving school)
- bet4fst: state, first exit from secondary school
- bet4info: documentation of construction of variable

## 3.5    Secondary school attendance after first exit [bex4….]

The majority of the population in Germany exits general school only once. A minority of the population returns to school after a period of working, attending post-secondary education,

---

[16] The variable bet4cert_g contains a few negative values. As negative values in SOEP data usually indicate missing values, the negative values have been recoded to positive values (-4=104, -3=103, -2=102, -1=101).

or engaging in other activities. These persons are likely to obtain (higher) school leaving certificates after they have left school for the first time. In the BIOEDU dataset we provide information on the year when we observe school attendance for the last time, including the type of school which was attended and the highest school leaving certificate ever obtained. For the majority of the population the information contained in these variables is the same as in the respective variables that cover the first exit from school (see Section 3.4).

The dataset contains the following variables on school attendance after first exit from school:

- bex4year: last observed year of school attendance
- bex4age: age (in years), last observed year of school attendance
- bex4type: type of school, last observed year of school attendance
- bex4cert: highest school leaving certificate ever observed
- bex4cert2: highest school leaving certificate ever observed abroad
- bex4fst: state, highest school leaving certificate

## 3.6       First entry into vocational training [bet5….]

Vocational training is the main option for those who have not obtained the *Abitur* (university entry qualification). Most holders of an *Abitur* degree do not choose to enroll in vocational training rather than a university degree program, but some do. In the majority of cases, the transition to vocational training takes place directly after secondary school ends or after military or civil service. Depending on the duration of school attendance, this is most likely to take place between the ages of 15 and 21, a period covered prospectively in the household questionnaire, the youth questionnaire, and the individual questionnaire. Hence, we use information on current vocational training participation from all three questionnaires to reconstruct transitions to vocational training. As in the case of exit from school, we provide detailed information on first entry and additional information on later transitions (see Section 3.8). In Germany, there is a distinction between full vocational training and pre-vocational training (*Berufsvorbereitung, Berufsorientierung*), which is also reflected in the SOEP questionnaires.  We do not consider participation in pre-vocational training to constitute entry into vocational training. Instead, we provide additional variables that record the first year of participation in such a course (and the age at that time, variables bet5z_year,

bet5z_age).[17] In the following, we describe the reconstruction of entries into full vocational training. In a first step, we reconstruct the transition based on the information on current vocational training participation. In a second step, we use additional information to determine the time of the transition precisely.

*Basic algorithm (Step 1)*

On the basis of yearly information on vocational training participation, we construct a variable of entry into training. We compare vocational training participation in t and t+1. If a person does not attend vocational training in t but in t+1, we assume that an entry into vocational training took place in the intervening period. As mentioned above, we consider only the first entries into vocational training. Using this simple algorithm, we observe an entry into vocational training for 5,661 cases.

*Additional checks (Step 1)*

1. *Earlier vocational training participation*: We deleted entries if they appeared not to be the first entries into vocational training (i.e., vocational training attendance observed in earlier years). This left 5,526 cases with an observed first entry into vocational training.

2. *Vocational degree or earlier drop-out*: We deleted entries if respondents had already obtained a vocational degree or reported a drop-out from vocational training at the time of the entry. This left 4,753 cases with an observed first entry into vocational training.

3. *Duration since exit from school*: For a few cases, we observed an entry into vocational training before the end of secondary school. These rather implausible cases were deleted, leaving 4,749 cases with an observed first entry into vocational training.

4. *Application of an age threshold*: We deleted cases that entered vocational training at an age of 14 years or younger. This left 4,738 cases with an observed first entry into vocational training.

5. *Earlier work in a qualified position*: We deleted cases for whom only one year of unspecified vocational training was observed (category "other") and who had previously worked in an occupation requiring a vocational degree. It is very likely that the persons with deleted entries attended a course of further training and did not enter

---

[17] There are a wide range of pre-vocational training courses available in Germany. Compared to other sources (Autorengruppe Bildungsberichterstattung 2010: 96) the SOEP contains a rather low percentage of individuals in such courses. As most of these courses are school-based, it is likely that respondents report school-based vocational training rather than pre-vocational training.

vocational training for the first time. This left 4,677 cases with an observed first entry into vocational training.

6. *Restoration of erroneously deleted exits*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible transitions. An example would be a person who, according to the data, attended vocational training at ages 13 and 16 and a general school in between. Using the basic algorithm, we would reconstruct two transitions: one between age 13 and 14, and a second between the ages of 15 and 16. Following our checks, we would delete the latter transition because the person appears to have participated in vocational training before this transition (see above). But we would also delete the former transition because of the age restriction. In this case, the reason for deleting the second assumed transition is no longer relevant. We therefore restored the information on the respective transition. After this last step, we observed the first entry into vocational training for 4,680 cases.

*Basic algorithm (Step 2)*

Although vocational training normally follows the typical academic calendar, the starting and ending dates of vocational training are not as standardized as those of primary and secondary school. For example, school leavers who have problems finding a suitable training position may enter vocational training at a later date. We therefore use additional information to precisely reconstruct the time when the transition took place. For firm-based vocational training, we use information on job changes or the date when a person started working for the current employer. If such information is not available, we use information from the activity calendar.

We use different sources in the following order:

1. *Starting date of new job since the beginning of the last year*: Every year, respondents are asked if they have started a new job since the beginning of the year before the survey. Since the majority of firm-based trainees consider themselves employed, we can use this variable as information on the start of vocational training.

2. *Starting date of work for current employer*: A second source is the information on when a person started to work for the current employer. If a person answers this question while in a spell of vocational training for the first time, we use this information to generate the entry date.

3. *Activity calendar data*: In some cases of firm-based training and in the majority of other cases, we cannot use the first two options to reconstruct the date of the exact transition. Hence, what we know from the data on educational participation is that a transition has taken place between wave t and wave t+1, a period which, in some cases, covers more than a year. In order to at least record the exact calendar year, we use additional information from the activity calendar (see Section 2.3). If, according to the activity calendar, a person was in training in the November of year t, we assume that the entry occurred in year t. If a person was not in training in November, we assume that the entry occurred in year t+1.

    a. *Firm-based training*: Firm-based vocational training is one category in the activity calendar (until 1999: together with further education). If a respondent was in firm-based training in November of year t, we assume that the transition took place in that year.

    b. *Other forms of vocational training*: The case is more difficult for other, school-based forms of vocational training. Basically, we follow the same approach but check more thoroughly whether or not the person was in general school or university (using data from the individual questionnaire).

We provide information on the source of the information on the date of entry into vocational training and the year when it was collected (variables bet5year_s, bet5year_g). In the majority of cases, the information is obtained from the question on job changes since the beginning of the year prior to the interview. In cases where we do not have any additional information on entry into vocational training, we assume that the entry took place in year t+1 (less than 10 percent).

The dataset contains the following variables on the first entry into vocational training:
- bet5obs: first entry into vocational training observed
- bet5year: first entry into vocational training, year
- bet5year_s: first entry into vocational training, source of exact year
- bet5year_g: gap between observed entry and source of exact year
- bet5month: first entry into vocational training, calendar month
- bet5age: first entry into vocational training, age in years
- bet5type: status before entry into vocational training
- bet5fst: : first entry into vocational training, state
- bet5info: entry into vocational training, construction of variable

- bet5z_year: pre-vocational training, first observed year
- bet5z_age: pre-vocational training, age at first observed year

## 3.7      First exit from vocational training [bet6….]

We reconstruct exits from vocational training in analogy to the reconstruction of entries (see Section 3.6 for details). Again, we consider exits from first-time training only. In a first step, we reconstruct the transition based on the information on current vocational training participation. In a second step, we add information on vocational degrees and the exact time of the transition (if available).

*Basic algorithm (Step 1)*

On the basis of yearly information on vocational training participation, we construct a variable of exit from training. We compare vocational training participation in t and t+1. If a person attended vocational training in t but not in t+1, we assume that an exit from vocational training took place in the intervening period. As already mentioned, we consider only the first exits from vocational training. Using this simple algorithm we observe an exit for 5,502 cases.

*Additional checks (Step 1)*

1.  *Deletion of double transitions*: Using the basic algorithm, we reconstructed multiple transitions for a number of cases (574 cases with 2 transitions, 45 cases with 3 transitions, 2 cases with 4 transitions). As our focus in this step was on first exits, we retained only the first observed exit. This did not change the number of cases for whom we observe an exit (n=5,502) but only the number of exits per case.

2.  *Vocational degree or earlier drop-out*: We deleted exits if respondents had already obtained a vocational degree or reported a drop-out from vocational training at the time of the exit. This left 4,438 cases with an observed first exit from vocational training.

3.  *Earlier work in a qualified position*: We deleted cases for whom we observed only one year of unspecified vocational training (category "other") and who had previously worked in an occupation that required a vocational degree. It is very likely that these persons attended further training and not first-time vocational training. This left 4,366 cases with an observed first entry into vocational training.

27

4.  *Duration since entry into vocational training*: For only a few cases did we observe an exit from vocational training prior to an entry. These rather implausible cases have been deleted, leaving 4,356 cases with an observed first exit from vocational training.

5.  *Application of an age threshold*: We deleted cases that exited vocational training at an age of 15 or younger. This left 4,347 cases with an observed first entry into vocational training.

6.  *Restoration of erroneously deleted exits*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible transitions. An example would be a person for whom we deleted one transition because it appeared to be the second but had also deleted the first based on the age threshold. In this case, the reason for deleting the second transition was no longer relevant. We therefore restored the information on the respective transition. After this last step, we reconstructed the transition to secondary school for 4,352 cases.

*Basic algorithm (Step 2)*

In Step 1 we reconstructed the exit from vocational training using information on educational participation. We also have data on degrees obtained and the exact months in which respondents completed different types of training (not for all cases). Generally, there are four different sources of information on educational degrees (see also Section 3.4):

-  Prospective data on completion of education (year prior to the survey): Every year respondents aged 17+ are asked if they have left vocational training since the beginning of the previous survey year and what type of degree they obtained (since 2000: if they dropped out without graduating).

-  Retrospective data on educational degrees: Usually in the first or second year of survey participation, new respondents complete the biographical questionnaire, which covers topics such as the respondent's social origins and life course up to entry into the SOEP. This includes questions on the highest educational degree obtained.

-  "Inventory" in 2000: The variables on the educational degrees are updated on a yearly basis if new educational degrees have been obtained (in the year prior to the survey). In addition, in the year 2000, every respondent in the sample was asked again about his or her highest educational degree. Unfortunately, educational degrees are not always consistent with information collected in earlier interviews.

- New respondents: From 1998 onwards, every first-time respondent has been asked about his or her highest educational degree as the information which is collected via the biographical questionnaire is usually not available in the first wave.

Our basic approach is to match the information on degrees to the year of the first exit from vocational training, which is reconstructed using information on educational participation (see Step 1). As described above, for many respondents, we also collect information on educational degrees retrospectively. This could be the same degree or a different one if a person participated in education between the two points of observation. In order to use the information on the degree that matches the first exit from vocational training, we use information collected at the time of the exit from training, which is, however, not always available. As a basic rule, we prioritize variables that contain precise information on the time when a given degree was obtained. We provide information on the source of the information on degrees and the year when it was collected (variables bet6cert_s bet6cert_g). In more than 85 percent of all cases, the information stems from the question on education completed since the beginning of the last year prior to the interview. In about a quarter of all cases, there is no information on degrees from vocational training. We cannot completely rule out the possibility that degrees were not recorded correctly in some of these cases. However, it is likely that the majority of these dropped out of vocational training before receiving a degree. If we look at the cases for which we have information on an entry (bet5obs=1) and an exit from vocational training, we see whether missing information is related to the duration in vocational training. In fact, the large majority of the respondents who attended vocational training for less than a year did not obtain a degree, while this share is not much higher than 10 percent after three or more years.

In the case that no information was collected on the month when a degree was obtained or when the respondent dropped out of training, we used additional information to precisely reconstruct the time when the transition took place. For firm-based vocational training, we use information on job changes or the date when a person finished working for the current employer. If such information is not available we use information from the so-called activity calendar (see Section 2.3).

The dataset contains the following variables on the first exit from vocational training:
- bet6obs: first exit from vocational training observed
- bet6year: first exit from vocational training, year
- bet6month: first exit from vocational training, calendar month
- bet6age: first exit from vocational training, age in years
- bet6type: first exit from vocational training, type of training
- bet6cert: first exit from vocational training, degree
- bet6cert_s: vocational degree, source
- bet6cert_g: gap between year of exit from vocational training and collection of data on degree
- bet6fst: : first exit from vocational training, state
- bet6info: first exit from vocational training, construction of variable

## 3.8 Vocational training after first exit [bex6….]

In addition to first entries to and exits from vocational training, we record the last year of vocational training participation and all types of vocational degrees ever obtained. For those who have not participated in vocational training since their first exit, the information contained in these variables is the same as in the respective variables that cover the first exit from vocational training (see Section 3.7).

The dataset contains the following variables on school attendance after first exit from school:
- bex6year: last observed year of vocational training
- bex6age: age (in years), last observed year of vocational training
- bex6type: type of school, last observed year of vocational training
- bex6cert1: observed vocational degree no 1, type
- bex6cert2: observed vocational degree no 2, type
- bex6cert3: observed vocational degree no 3, type
- bex6cert4: observed vocational degree no 4, type
- bex6certn: number of observed vocational degrees
- bex6fst: state, last observed year of vocational training

## 3.9 First entry into tertiary education [bet7….]

More than 75 percent of the students who obtained the *Abitur* qualification at the end of school pursue tertiary education (see Autorengruppe Bildungsberichterstattung 2010: 118). Holders of other school leaving certificates are not allowed to enroll in the university, with only a few exceptions.[18] Transitions to university usually do not occur before the age of 18. Hence, participation in tertiary education is covered in the individual questionnaire only (and not in the household or youth questionnaire). As in the case of the other transitions, we provide detailed information on first entry and additional information on later transitions (see Section 3.11). In a first step, we reconstruct the transition based on the information on current university (or university of applied sciences) attendance. In a second step, we use additional information to determine the time of the transition precisely.

*Basic algorithm (Step 1)*

On the basis of yearly information on tertiary education participation we construct a variable of entry into tertiary education. We compare tertiary education participation in t and t+1. If a person is not attending university in t but in t+1, we assume that an entry into tertiary education took place in the intervening period. As already mentioned, we consider only the first entries into tertiary education. By using this simple algorithm we observe an entry for 2,314 cases.

*Additional checks (Step 1)*

1. *Earlier tertiary education participation*: We deleted entries if they appeared not to be the first entries into tertiary education (i.e., attendance of a university or university of applied sciences in earlier years). This left 2,211 cases with an observed first entry into tertiary education.

2. *Deletion of double transitions*: Using the basic algorithm, we reconstructed multiple transitions for a number of cases (196 cases with 2 transitions, 15 cases with 3 transitions, 4 cases with 4 or more transitions). As our focus in this step was on first entry, we retained only the first observed entry. This did not change the number of cases for whom we observe an entry (n=2,211) but only the number of entries per case.

3. *Tertiary degree or earlier drop-out*: We deleted entries if respondents had already obtained a university (or university of applied sciences) degree or reported a drop-

---

[18] Entry into universities of applied sciences and some other universities is possible after completing vocational training and obtaining work experience. In some exceptional cases, "genius clauses" may apply. Entry into art and music schools usually does not require the Abitur exam.

out from tertiary education at the time of the entry. This left 2,041 cases with an observed first entry into tertiary education.

4. *Duration since exit from school*: For a few cases, we observed entry into tertiary education before the end of secondary school. These rather implausible transitions were deleted, leaving 2,040 cases with an observed first entry into tertiary education.

5. *Restoring erroneously deleted entries*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible transitions. An example would be a person for whom we deleted a transition because it appeared to be the second but had also deleted the first based on the duration threshold. In this case, the reason for deleting the second transition was no longer relevant. We therefore restored the information on the respective transition. After this last step we reconstructed the transition to secondary school for 2,041 cases.

*Basic algorithm (Step 2)*

Using the basic algorithm and the additional checks, we were able to reconstruct transitions that took place between two panel waves, a time period of around 12 months on average. In the case of entries into vocational training, we used additional information to determine the exact month of the transition. However, in the case of entries into tertiary education, additional information is scarce. As a rule, the year of the transition was determined in the year tertiary education was first observed. That is, when a respondent reported in an April 2010 interview that she was attending a university, the year of the transition was defined as 2010. However, the transition could have taken place the year before if the last interview was in April 2009 and she had started in October 2009. Therefore, we checked in the activity calendar data (see Section 2.3) to see if a person had participated in tertiary education in the November prior to the year of the interview when tertiary participation was first observed in the personal interview. In these cases, we defined the year of the transition as the year prior to the interview (2009 instead of 2010 in our example). Information on the year of the transition is contained in the variable bet7year_s.

The dataset contains the following variables on the first entry into tertiary education:
- bet7obs: first entry into tertiary education observed
- bet7year: first entry into tertiary education, year
- bet7year_s: first entry into tertiary education, source of exact year
- bet7age: first entry into tertiary education, age in years

- bet7type: status before entry into tertiary education
- bet7info: entry tertiary education, construction of variable

## 3.10 First exit from tertiary education [bet8….]

We reconstructed the exit from tertiary education in analogy to the reconstruction of the entries (see Section 3.9 for details). Again, we considered exits from first-time training only. In a first step, we reconstructed the transition based on the information on current tertiary education participation. In a second step, we used additional information to locate the year of the transition.

*Basic algorithm (Step 1)*

On the basis of yearly information on tertiary education participation we constructed a variable of exit from university. We compared tertiary education participation in t and t+1. If a person attended a university (or university of applied sciences) in t but not in t+1, we assumed that an exit from tertiary education took place in the intervening period.[19] As already mentioned, we consider only the first exits from tertiary education. Using this simple algorithm, we observed an exit for 2,041 cases.

*Additional checks (Step 1)*

1. *Deletion of double transitions*: Using the basic algorithm, we reconstructed multiple transitions for a number of cases (228 cases with 2 transitions, 19 cases with 3 transitions, 6 cases with 4 or more transitions). As our focus in this step was on first exits, we retained only the first observed exit. This did not change the number of cases for whom we observed an exit (n=2,041) but only the number of exits per case.

2. *Tertiary degree or earlier drop-out*: We deleted exits if respondents had already obtained a tertiary degree or reported a drop-out from university at the time of the exit. This left 1,738 cases with an observed first exit from tertiary education.

3. *Duration since entry into vocational training*: For a few cases we observed exits from tertiary education earlier than an entry. These rather implausible transitions were deleted, leaving 1,725 cases with an observed first exit from tertiary education.

4. *Restoration of erroneously deleted exits*: In some cases, we observed two transitions using the basic algorithm, both of which were deleted when checking for implausible

---

[19] In a few cases we used information from the "gap"-datasets ($PLUECKE).

transitions. An example would be a person for whom we deleted a transition because it appeared to be the second but had deleted also the first based on the age threshold. In this case, the reason for deleting the second transition is no longer relevant. We therefore restored the information on the respective transition. After this last step, we reconstructed the transition to secondary school for 1,733 cases.

*Basic algorithm (Step 2)*

In Step 1 we reconstructed the exit from tertiary education using information on educational participation. We also have data on degrees obtained and the exact month of the end of tertiary education participation (although not for all cases). Generally, there are four different sources of information on educational degrees (see also Section 3.4):

- Prospectively collected data on completion of education (year prior to the survey): Every year, all respondents aged 17+ years are asked whether they left tertiary education since the beginning of the previous survey year and what type of degrees they obtained (since 2000: whether they dropped out without graduating).
- Retrospective data on educational degrees: Usually in the first or second year of survey participation, new respondents complete the biographical questionnaire, which covers topics such as the respondent's social origins and life course up to entry into the SOEP. This includes questions on the highest educational degree obtained.
- "Inventory" in 2000: The variables on educational degrees are updated on a yearly basis if new educational degrees have been obtained (in the year prior to the survey). In addition, in the year 2000, every respondent in the sample was asked again about his or her highest educational degree. Unfortunately, educational degrees are not always consistent with the information collected in earlier interviews.
- New respondents: Since 1998, every first-time respondent has been asked about his or her highest educational degree as the information collected in the biographical questionnaire is usually not available in the first wave.

Our basic approach is to match the information on degrees to the year of the first exit from tertiary education, which is reconstructed using information on educational participation (see Step 1). As described above, for many respondents, we collect information on educa-

tional degrees retrospectively as well. This information could refer the same degree or a different one if a person participated in education between the two points of observation. In order to use the information on the degree that matches the first exit from tertiary education, we use information collected at the time of the exit from education, which is, however, not always available. As a basic rule, we prioritize variables that contain precise information on the time when a given degree was obtained. We provide information on the source of the information on degrees and the year when it has been collected (variables bet8cert_s bet8cert_g). In more than 90 percent of the cases the information stems from the question on leaving tertiary education since the beginning of the last year prior to the interview. In more than a third of all cases there is no information on degrees from tertiary education. We cannot completely rule out that for some of these cases an obtained degree was not correctly recorded. However, it is most likely that the majority of these cases dropped out of tertiary education and have not obtained a degree. If we look at the cases for which we have information on an entry (bet7obs=1) and an exit from tertiary education we can see if missing information is related to the duration in tertiary education. In fact, the large majority of the respondents who attended tertiary education for less than a three years did not obtain a degree while this share is below 15 percent after five and more years.

In the case that no information on the month when a degree was obtained or when a drop-out took place was collected we used additional information from the activity calendar (see Section 2.3) to precisely reconstruct the time when the transition took place. Information on how the year of the end of tertiary education was determined is stored in the variable bet8year_s.

The dataset contains the following variables on the first exit from vocational training:
- bet8year: first exit from tertiary, year
- bet8year_s: first exit from tertiary, source of year
- bet8month: first exit from tertiary, calendar month
- bet8age: first exit from tertiary, age in years
- bet8type: first exit from tertiary, type of school
- bet8cert: first exit from tertiary, degree
- bet8cert_s: tertiary degree, source
- bet8cert_g: gap between year of exit from tertiary and collection of data on degree
- bet8fst: : first exit from tertiary, state
- bet8info: first exit from tertiary, construction of variable

## 3.11 Tertiary education after first exit [bex8....]

In addition to first entries to and exits from the university, we record the last year of tertiary education participation and all tertiary degrees ever obtained. For those who have not participated in tertiary education since their first exit, the information contained in these variables is the same as in the respective variables that cover the first exit from university (see Section 3.10).

The dataset contains the following variables on school attendance after first exit from school:

- bex8year: last observed year in tertiary education
- bex6age: age (in years), last observed year in tertiary education
- bex8type: type of school, last observed year in tertiary education
- bex8cert: tertiary degree
- bex8cert2: tertiary degree (abroad)
- bex8fst: state, last observed year in tertiary education

## 3.12 Variables documenting the generation of educational transitions

Due to temporary drop-outs, item non-response, unequal spacing of panel waves, misclassification, etc. we cannot always unambiguously reconstruct if and exactly when a transition took place. Therefore, we provide variables (bet2info-bet8info) that record ambiguities in the data that may affect the reconstruction of the transitions. The variables have nine digits that can take the values 0 or 1. Digits 1-4 refer to cases where we initially have reconstructed more than one transition and deleted one or more of those following the algorithms described in Section 3). Digit 5 refers to cases that were not considered for transition in/out of vocational training/tertiary education because the respondents reported already holding a degree at the time of the transition (i.e., not first transition). Digits 6-8 provide additional information on some minor ambiguities. Digit 9 refers to additional criteria used in the reconstruction of the transitions. For more detailed information, see the sections that describe the reconstruction of the respective transitions and Appendix II for frequency tables. In short, the single digits of the variables betXinfo refer to the following steps in the process of data reconstruction:

1. Following the basic algorithm (t compared to t+1), we reconstructed two or more transitions. Since we expect only one transition (t2, t3) or consider only the first, we deleted additional transitions (t4-t8).

2. Transitions were deleted if the duration since the earlier transition seemed unreliably short (t3-t8).

3. Transitions were deleted if the age at the transition was implausible (t2-t8).

4. Following various criteria, we deleted all initially reconstructed transitions. In a further step, we chose the most plausible one (t2-t8).

5. Cases not considered for transitions to/from vocational training/tertiary education because the respondents reported already holding a degree at the time of the transition (i.e., not first transition).

6. Minor inconsistencies in the observed participation in education.

7. Unlikely parallel activities during participation in education.

8. Problems in the exact timing of the transition.

9. Refers to Lower Saxony and Bavaria, where specific regulations (*Orientierungsstufe, Teilhauptschule*) that are not fully reflected in the questionnaire make the reconstruction of the transition to secondary school less reliable (t3, see Section 3.3). Refers also to cases which, according to the data, did not obtain university entry qualifications but participated in tertiary education (t7/t8, see Sections 3.9 and 3.10).

# 4 Some final remarks on using the dataset

When using this dataset, keep in mind that most of the information covered was not obtained directly from questions in the SOEP questionnaires but derived from combinations of several variables. In the process of reconstruction, assumptions have been made that we have tried to describe in the previous sections in as detailed a manner as possible. The more additional information we were able to use as the basis for our assumptions, e.g., strict institutional regulations (on compulsory schooling, etc.), the more accurately we were able to reconstruct the transitions.

The dataset covers transitions starting in early childhood up to tertiary education. For part of the sample, only one of these transitions or episodes is observed; for others the whole sequence from primary education up to exits from tertiary education. The variable *beinfo* provides an overview of the frequencies of these different patterns (see Table 1). In total, the dataset contains information on 57,250 persons. This is the part of the SOEP sample for

whom we have observed an educational transition and/or an educational degree. We have full information for 389 cases (pattern 811111111).

We have provided a number of variables for which we documented the process of data generation and the sources of the data (betXinfo, variables with suffixes _s or _g). These variables can be used as indicators of the degree of uncertainty in the process of the reconstruction of educational transitions. The degree of uncertainty is higher where the variables could not be reconstructed using just the basic algorithm (e.g., bet2info<>"0000|0|0000"). The same applies to long durations between an observed exit and the observation of a matching educational degree (e.g., a high value in bet6cert_g). It is certainly advisable to check whether certain deviations in the process of data generation "explain" substantial results. For example, if children living in households who where interviewed in August (this information is provided in betXinfo) have a much higher propensity of starting school late (bet2agemo), this might just be a data artifact because it is difficult to decide whether the information the household provided referred to the school year that just started in August or to the school year that just ended at the time of the interview. In general, one can expect that there are no such systematic measurement errors in the reconstructed variables. However, if you want to take a closer look at potential biases, you can use the variables that document the data generation process. This documentation describes a beta version of the dataset (v27_0.2). If you have any comments or encounter any problems in using the dataset, please let us know.

# References

*Autorengruppe Bildungsberichterstattung* (2010): Bildungsbericht 2010. Ein indikatorengestützter Bericht mit einer Analyse zu den Analysen des Bildungswesens im demografischen Wandel. Bielefeld: Bertelsmann.

*Frick*, Joachim R./ *Lohmann*, Henning (2010): Biography and Life History Data in the German Socio Economic Panel (SOEP, v26, 1984-2009). DIW Data Documentation 52. Berlin.

*Kultusministerkonferenz* – KMK (1994): Schüler, Klassen, Lehrer und Absolventen der Schulen 1984 bis 1993. Statistische Veröffentlichungen der Kultusministerkonferenz. Dokumentation Nr. 129. Bonn.

*Kultusministerkonferenz* – KMK (div. years): Schüler, Klassen, Lehrer und Absolventen der Schulen. Download: www.kmk.org/statistik/schule/statistiken/schueler-klassen-lehrer-und-absolventen-der-schulen.html (20.07.09).

*Lohmann*, Henning/ *Spieß*, C. Katharina/ *Groh-Samberg*, Olaf (2009): Analysepotenziale des Sozio-oekonomischen Panels (SOEP) für die empirische Bildungsforschung. In: Zeitschrift für Erziehungswissenschaft 12(2): 252-280.

## Tables

Table 1:
**Overview on observed patterns (Variable `beinfo`)**

```
pattern of |
observation |        Freq.      Percent         Cum.
------------+----------------------------------
        -2 |           50         0.09         0.09
  100000001 |       41,272        72.09        72.18
  100000010 |           15         0.03        72.20
  100000100 |           16         0.03        72.23
  100001000 |           54         0.09        72.33
  100010000 |          922         1.61        73.94
  100100000 |          109         0.19        74.13
  101000000 |          142         0.25        74.38
  110000000 |        1,661         2.90        77.28
  200000011 |        1,455         2.54        79.82
  200000101 |          209         0.37        80.18
  200000110 |            6         0.01        80.19
  200001001 |          192         0.34        80.53
  200001010 |            1         0.00        80.53
  200001100 |          146         0.26        80.79
  200010001 |           75         0.13        80.92
  200010100 |            7         0.01        80.93
  200011000 |           16         0.03        80.96
  200110000 |           56         0.10        81.06
  201000001 |            4         0.01        81.06
  201001000 |            2         0.00        81.07
  201010000 |          101         0.18        81.24
  201100000 |          457         0.80        82.04
  210010000 |            5         0.01        82.05
  210100000 |          198         0.35        82.39
  211000000 |          197         0.34        82.74
  300000111 |          707         1.23        83.97
  300001011 |            3         0.01        83.98
  300001101 |          813         1.42        85.40
  300001110 |           58         0.10        85.50
  300010011 |           23         0.04        85.54
  300010101 |           12         0.02        85.56
  300011001 |           62         0.11        85.67
  300011100 |           33         0.06        85.73
  300101001 |            1         0.00        85.73
  300101100 |            1         0.00        85.73
  300110001 |            3         0.01        85.74
  300110100 |            1         0.00        85.74
  301000011 |            2         0.00        85.74
  301000110 |            1         0.00        85.74
  301001100 |            1         0.00        85.74
  301010001 |            4         0.01        85.75
  301010100 |            1         0.00        85.75
  301011000 |            2         0.00        85.76
  301100001 |            5         0.01        85.77
  301101000 |            2         0.00        85.77
  301110000 |          570         1.00        86.77
  310110000 |           13         0.02        86.79
```

| | | | |
|---|---|---|---|
| 311000001 | 6 | 0.01 | 86.80 |
| 311001000 | 1 | 0.00 | 86.80 |
| 311010000 | 119 | 0.21 | 87.01 |
| 311100000 | 1,349 | 2.36 | 89.36 |
| 400001111 | 1,979 | 3.46 | 92.82 |
| 400010111 | 16 | 0.03 | 92.85 |
| 400011011 | 3 | 0.01 | 92.85 |
| 400011101 | 200 | 0.35 | 93.20 |
| 400011110 | 13 | 0.02 | 93.23 |
| 400101101 | 2 | 0.00 | 93.23 |
| 400110011 | 2 | 0.00 | 93.23 |
| 400111001 | 4 | 0.01 | 93.24 |
| 400111100 | 1 | 0.00 | 93.24 |
| 401000111 | 3 | 0.01 | 93.25 |
| 401001101 | 18 | 0.03 | 93.28 |
| 401001110 | 2 | 0.00 | 93.28 |
| 401010011 | 2 | 0.00 | 93.29 |
| 401010101 | 2 | 0.00 | 93.29 |
| 401011001 | 3 | 0.01 | 93.29 |
| 401011100 | 3 | 0.01 | 93.30 |
| 401101001 | 3 | 0.01 | 93.30 |
| 401101100 | 4 | 0.01 | 93.31 |
| 401110001 | 31 | 0.05 | 93.37 |
| 401110100 | 2 | 0.00 | 93.37 |
| 401111000 | 5 | 0.01 | 93.38 |
| 411000011 | 3 | 0.01 | 93.38 |
| 411001001 | 2 | 0.00 | 93.39 |
| 411010001 | 8 | 0.01 | 93.40 |
| 411011000 | 3 | 0.01 | 93.41 |
| 411100001 | 8 | 0.01 | 93.42 |
| 411101000 | 4 | 0.01 | 93.43 |
| 411110000 | 1,402 | 2.45 | 95.88 |
| 500011111 | 460 | 0.80 | 96.68 |
| 500101111 | 4 | 0.01 | 96.69 |
| 500110111 | 2 | 0.00 | 96.69 |
| 500111011 | 1 | 0.00 | 96.69 |
| 500111101 | 16 | 0.03 | 96.72 |
| 500111110 | 2 | 0.00 | 96.72 |
| 501001111 | 52 | 0.09 | 96.81 |
| 501010111 | 1 | 0.00 | 96.82 |
| 501011101 | 24 | 0.04 | 96.86 |
| 501011110 | 2 | 0.00 | 96.86 |
| 501100111 | 1 | 0.00 | 96.86 |
| 501101101 | 15 | 0.03 | 96.89 |
| 501101110 | 1 | 0.00 | 96.89 |
| 501110011 | 4 | 0.01 | 96.90 |
| 501110101 | 1 | 0.00 | 96.90 |
| 501110110 | 1 | 0.00 | 96.90 |
| 501111001 | 21 | 0.04 | 96.94 |
| 501111100 | 21 | 0.04 | 96.97 |
| 511001101 | 9 | 0.02 | 96.99 |
| 511001110 | 1 | 0.00 | 96.99 |
| 511010011 | 3 | 0.01 | 97.00 |
| 511011001 | 4 | 0.01 | 97.00 |
| 511011100 | 5 | 0.01 | 97.01 |
| 511100011 | 2 | 0.00 | 97.02 |
| 511100110 | 1 | 0.00 | 97.02 |
| 511101001 | 11 | 0.02 | 97.04 |
| 511101100 | 14 | 0.02 | 97.06 |
| 511110001 | 52 | 0.09 | 97.15 |

```
511110010 |          4           0.01          97.16
511110100 |          3           0.01          97.17
511111000 |         16           0.03          97.19
600111111 |         31           0.05          97.25
601011111 |         76           0.13          97.38
601101111 |         34           0.06          97.44
601110111 |         13           0.02          97.46
601111011 |          1           0.00          97.46
601111101 |        130           0.23          97.69
601111110 |          6           0.01          97.70
611001111 |         18           0.03          97.73
611010111 |          2           0.00          97.74
611011101 |         22           0.04          97.77
611011110 |          3           0.01          97.78
611100111 |          1           0.00          97.78
611101101 |         35           0.06          97.84
611101110 |          3           0.01          97.85
611110011 |         16           0.03          97.88
611110101 |          9           0.02          97.89
611111001 |         67           0.12          98.01
611111100 |         71           0.12          98.13
701111111 |        299           0.52          98.66
711011111 |         37           0.06          98.72
711101111 |         33           0.06          98.78
711110111 |         12           0.02          98.80
711111011 |          1           0.00          98.80
711111101 |        278           0.49          99.29
711111110 |         20           0.03          99.32
811111111 |        389           0.68         100.00
-----------+-------------------------------
    Total |     57,250         100.00
```

Note: 1. digit: number of observed transitions/episodes, 2.: first year of ECEC, 3.: last year of ECEC, 4. school enrollment, 5. transition to secondary school, 6. exit from secondary school, 7. transition to vocational training/tertiary education or observed until age 23, 8. transition from vocational training/tertiary education or observed until age 27, 9. highest school leaving certificate.

# Appendix

## Appendix I: List of variables

*Identifiers and meta data*

persnr         permanent individual ID number
firstobs       age at first observation
lastobs        age at last observation
gebjahr        year of birth
gebmonat       month of birth
beinfo         pattern of observation

*Early childhood education and care (ECEC)*

bet0age        age (y.), first observed attendance of ECEC
bet0agem       age (mon.), first observed attendance of ECEC
bet0fst        state, first observed attendance of ECEC
bet0obs        uncensored first attendance of ECEC
bet0year       Year, first observed attendance of ECEC
bet1age        age (year), last observed attendance of ECEC
bet1agem       age (month), last observed attendance of ECEC
bet1fst        state, last observed attendance of ECEC
bet1obs        uncensored last attendance of ECEC
bet1year       Year, last observed attendance of ECEC

*Primary education*

bet2age        school enrollment, age in years
bet2agem       school enrollment, age in months
bet2fst        school enrollment, state
bet2info       school enrollment, construction of variable
bet2obs        school enrollment observed
bet2type       school enrollment, type of school
bet2year       school enrollment, year

*Transition to secondary school*

bet3age        transition to secondary school, age in years
bet3agem       transition to secondary school, age in months
bet3fst        transition to secondary school, state
bet3info       transition to secondary school, construction of variable
bet3obs        transition to secondary school observed
bet3type       transition to secondary school, type of school
bet3year       transition to secondary school, year

---

*First exit from secondary school*

| | |
|---|---|
| bet4age | first exit from school, age in years |
| bet4cert | first exit from school, certificate |
| bet4cert_g | gap between year of exit from school and collection of data on school-leaving certificate |
| bet4cert_s | school-leaving certificate, source |
| bet4fst | first exit from school, state |
| bet4info | school exit, construction of variable |
| bet4obs | first exit from school observed |
| bet4type | first exit from school, type of school |
| bet4year | first exit from school, year |

*First entry into vocational training*

| | |
|---|---|
| bet5age | first entry into vocational training, age in years |
| bet5fst | first entry into vocational training, state |
| bet5info | entry vocational, construction of variable |
| bet5month | first entry into vocational training, calendar month |
| bet5obs | first entry into vocational training observed |
| bet5type | first entry into vocational training, type of training |
| bet5year | first entry into vocational training, year |
| bet5year_g | gap between observed entry and source of year |
| bet5year_s | first entry into vocational training, source of year |
| bet5z_age | pre-vocational training, age at first observed year |
| bet5z_year | pre-vocational training, first observed year |

*First exit from vocational training*

| | |
|---|---|
| bet6age | first exit from vocational training, age in years |
| bet6cert | first exit from vocational training, degree |
| bet6cert_g | gap between year of exit from vocational training and collection of data on degree |
| bet6cert_s | vocational degree, source |
| bet6fst | first exit from vocational training, state |
| bet6info | exit vocational training, construction of variable |
| bet6month | first exit from vocational training, calendar month |
| bet6obs | first exit from vocational training observed |
| bet6type | first exit from vocational training, type of training |
| bet6year | first exit from vocational training, year |

*First entry into tertiary education*

| | |
|---|---|
| bet7age | first entry into tertiary education, age in years |
| bet7fst | first entry into tertiary education, state |
| bet7info | first entry into tertiary education, construction of variable |
| bet7obs | first entry into tertiary education observed |
| bet7type | first entry into tertiary education, type of school |

44

bet7year    first entry into tertiary education, year
bet7year_s  first entry into tertiary education, source of year


*First exit from tertiary education*
bet8age     first exit from tertiary education, age in years
bet8cert     first exit from tertiary education, degree
bet8cert_g  gap between year of exit from tertiary education and collection of data on degree
bet8cert_s  tertiary degree, source
bet8fst      first exit from tertiary education, state
bet8info     first exit from tertiary education, construction of variable
bet8month  first exit from tertiary education, calendar month
bet8obs     first exit from tertiary education observed
bet8type    first exit from tertiary education, type of training
bet8year    first exit from tertiary education, year
bet8year_s  first exit from tertiary education, source of year


*Highest school leaving certificate and last observed year in school*
bex4age     last observed year in school, age in years
bex4cert    highest school-leaving certificate obtained
bex4cert2  highest school-leaving certificate (abroad)
bex4fst     last observed year in school, state
bex4type    last observed year in school, type of school
bex4year    last observed year in school, year


*Vocational degree and last observed year in vocational training*
bex6age     last observed year in vocational training, age in years
bex6cert1  observed vocational degree no. 1, type
bex6cert2  observed vocational degree no. 2, type
bex6cert3  observed vocational degree no. 3, type
bex6cert4  observed vocational degree abroad, type
bex6certn  observed vocational degrees, number
bex6fst     last observed year in vocational training, state
bex6type    last observed year in vocational training, type of school
bex6year    last observed year in vocational training, year


*Tertiary degree and last observed year in tertiary education*
bex8age     last observed year in tertiary education, age in years
bex8cert    tertiary degree
bex8cert2  tertiary degree (abroad)
bex8fst     last observed year in tertiary education, state
bex8type    last observed year in tertiary education, type of school
bex8year    last observed year in tertiary education, year

## Appendix II: Variables with information on the reconstruction of transitions (betXinfo)

```
-> tabulation of bet2info

     school |
  enrolment, |
  constructio |
        n of |
     variable |       Freq.       Percent         Cum.
------------+-------------------------------------------
         -2 |      51,283         89.58        89.58
0000|0|0000 |       5,533          9.66        99.24
0000|0|0010 |         253          0.44        99.68
0010|0|0000 |          55          0.10        99.78
1000|0|0000 |          96          0.17        99.95
1000|0|0010 |           2          0.00        99.95
1010|0|0000 |           2          0.00        99.95
1011|0|0000 |          26          0.05       100.00
------------+-------------------------------------------
      Total |      57,250        100.00

-> tabulation of bet3info

  transition |
     to sec. |
      school, |
  constructio |
        n of |
     variable |       Freq.       Percent         Cum.
------------+-------------------------------------------
         -2 |      51,100         89.26        89.26
0000|0|0000 |       5,053          8.83        98.08
0000|0|0001 |         119          0.21        98.29
0000|0|0010 |         274          0.48        98.77
0000|0|0011 |           7          0.01        98.78
0010|0|0000 |          96          0.17        98.95
0100|0|0000 |         174          0.30        99.25
1000|0|0000 |         381          0.67        99.92
1000|0|0001 |           7          0.01        99.93
1000|0|0010 |          12          0.02        99.95
1010|0|0000 |           3          0.01        99.96
1011|0|0000 |          11          0.02        99.98
1100|0|0000 |           3          0.01        99.98
1101|0|0000 |          10          0.02       100.00
------------+-------------------------------------------
      Total |      57,250        100.00
```

---

```
-> tabulation of bet4info
```

| school exit, constructio n of variable | Freq. | Percent | Cum. |
|---|---|---|---|
| -2 | 51,137 | 89.32 | 89.32 |
| 0000\|0\|0000 | 5,131 | 8.96 | 98.28 |
| 0000\|0\|0010 | 257 | 0.45 | 98.73 |
| 0000\|0\|1000 | 143 | 0.25 | 98.98 |
| 0000\|0\|1010 | 13 | 0.02 | 99.01 |
| 0010\|0\|0000 | 84 | 0.15 | 99.15 |
| 0100\|0\|0000 | 17 | 0.03 | 99.18 |
| 1000\|0\|0000 | 403 | 0.70 | 99.89 |
| 1000\|0\|0010 | 9 | 0.02 | 99.90 |
| 1000\|0\|1000 | 3 | 0.01 | 99.91 |
| 1001\|0\|0000 | 28 | 0.05 | 99.96 |
| 1001\|0\|0010 | 2 | 0.00 | 99.96 |
| 1001\|0\|1000 | 5 | 0.01 | 99.97 |
| 1010\|0\|0000 | 3 | 0.01 | 99.97 |
| 1011\|0\|0000 | 11 | 0.02 | 99.99 |
| 1011\|0\|1000 | 1 | 0.00 | 99.99 |
| 1100\|0\|0000 | 2 | 0.00 | 100.00 |
| 1101\|0\|0000 | 1 | 0.00 | 100.00 |
| Total | 57,250 | 100.00 | |

```
-> tabulation of bet5info
```

| entry vocational, constructio n of variable | Freq. | Percent | Cum. |
|---|---|---|---|
| -2 | 51,589 | 90.11 | 90.11 |
| 0000\|0\|0000 | 3,974 | 6.94 | 97.05 |
| 0000\|0\|0010 | 37 | 0.06 | 97.12 |
| 0000\|0\|0100 | 2 | 0.00 | 97.12 |
| 0000\|0\|1000 | 112 | 0.20 | 97.32 |
| 0000\|0\|1010 | 2 | 0.00 | 97.32 |
| 0000\|0\|1100 | 1 | 0.00 | 97.32 |
| 0000\|1\|0000 | 703 | 1.23 | 98.55 |
| 0010\|0\|0000 | 2 | 0.00 | 98.55 |
| 0100\|0\|0000 | 3 | 0.01 | 98.56 |
| 1000\|0\|0000 | 648 | 1.13 | 99.69 |
| 1000\|0\|0010 | 11 | 0.02 | 99.71 |
| 1000\|0\|0100 | 2 | 0.00 | 99.71 |
| 1000\|0\|1000 | 80 | 0.14 | 99.85 |
| 1000\|0\|1100 | 5 | 0.01 | 99.86 |
| 1000\|1\|0000 | 70 | 0.12 | 99.98 |
| 1001\|0\|0000 | 3 | 0.01 | 99.99 |
| 1100\|0\|0000 | 6 | 0.01 | 100.00 |
| Total | 57,250 | 100.00 | |

```
-> tabulation of bet6info

     exit |
vocational, |
constructio |
     n of |
  variable |      Freq.      Percent       Cum.
------------+-----------------------------------
        -2 |     51,758        90.41       90.41
0000|0|0000 |      3,557         6.21       96.62
0000|0|1000 |        288         0.50       97.12
0000|1|0000 |        954         1.67       98.79
0010|0|0000 |          4         0.01       98.80
0100|0|0000 |          7         0.01       98.81
1000|0|0000 |        547         0.96       99.76
1000|0|1000 |         22         0.04       99.80
1000|1|0000 |        105         0.18       99.99
1100|0|0000 |          8         0.01      100.00
------------+-----------------------------------
     Total |     57,250       100.00

-> tabulation of bet7info

    entry |
 tertiary, |
constructio |
     n of |
  variable |      Freq.      Percent       Cum.
------------+-----------------------------------
        -2 |     54,937        95.96       95.96
0000|0|0000 |      1,587         2.77       98.73
0000|0|0001 |        205         0.36       99.09
0000|0|0100 |         32         0.06       99.15
0000|0|0101 |         11         0.02       99.17
0000|0|1000 |          4         0.01       99.17
0000|1|0000 |        157         0.27       99.45
1000|0|0000 |        279         0.49       99.93
1000|0|0001 |         11         0.02       99.95
1000|0|0100 |          8         0.01       99.97
1000|0|0101 |          2         0.00       99.97
1000|0|1000 |          4         0.01       99.98
1000|1|0000 |         12         0.02      100.00
1101|0|0000 |          1         0.00      100.00
------------+-----------------------------------
     Total |     57,250       100.00
```

```
-> tabulation of bet8info

     exit |
  tertiary, |
constructio |
      n of |
   variable |     Freq.      Percent       Cum.
------------+--------------------------------
        -2 |    55,209        96.43        96.43
0000|0|0000 |     1,236         2.16        98.59
0000|0|0001 |       180         0.31        98.91
0000|0|0100 |        27         0.05        98.96
0000|0|0101 |        14         0.02        98.98
0000|0|1000 |        56         0.10        99.08
0000|0|1001 |         4         0.01        99.08
0000|0|1100 |         1         0.00        99.09
0000|0|1101 |         1         0.00        99.09
0000|1|0000 |       266         0.46        99.55
0100|0|0000 |         3         0.01        99.56
1000|0|0000 |        86         0.15        99.71
1000|0|0001 |        12         0.02        99.73
1000|0|0100 |         5         0.01        99.74
1000|0|1000 |        96         0.17        99.91
1000|0|1001 |         4         0.01        99.91
1000|0|1100 |         2         0.00        99.92
1000|0|1101 |         1         0.00        99.92
1000|1|0000 |        37         0.06        99.98
1100|0|0000 |         2         0.00        99.99
1101|0|0000 |         4         0.01        99.99
1101|0|0001 |         2         0.00       100.00
1101|0|1000 |         2         0.00       100.00
------------+--------------------------------
      Total |    57,250       100.00
```