Discussion Paper No. 16-020

Imputation Rules for the Implementation of the Pre-Unication Education Variable in the BASiD Data Set

Nicole Gürtzgen and André Nolte

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 16-020

Imputation Rules for the Implementation of the Pre-Unication Education Variable in the BASiD Data Set

Nicole Gürtzgen and André Nolte

Imputation Rules for the Implementation of the Pre-Unification Education Variable in the *BASiD* Data Set

Nicole Gürtzgen[1, 2)] and André Nolte[3)]

[1)] Institute for Employment Research, Nuremberg

[2)] University of Regensburg

[3)] Centre for European Economic Research, Mannheim*

March 10, 2016

Abstract

Using combined data from the German Pension Insurance and the Federal Employment Agency (*BASiD*), this study proposes different procedures for imputing the pre-unification education variable in the *BASiD* data. To do so, we exploit information on education-related periods that are creditable for the Pension Insurance. Combining these periods with information on the educational system in the former GDR, we propose three different imputation procedures, which we validate using external GDR census data for selected age groups in 1981. A common result from all procedures is that they tend to underpredict (overpredict) the share of high-skilled (low-skilled) for the oldest age groups. Comparing our imputed education variable with information on educational attainment from the Integrated Employment Biographies (*IEB*) reveals that the best match is obtained for the vocational training degree. Although regressions show that misclassification with respect to *IEB* information is clearly related to observables, we do not find any systematic pattern across skill groups.

JEL Classification: I2, C81
Keywords: Imputation Rules, Administrative Data, East Germany, Education, Institutions

*Full address of correspondence: Nicole Gürtzgen, Institute for Employment Research, Regensburger Str. 104, D-90478 Nuremberg, E-Mail: nicole.guertzgen@iab.de; André Nolte, Centre for European Economic Research, Department of Labour Markets, Human Resources and Social Policy, L 7.1, D-68161 Mannheim, E-Mail: nolte@zew.de. We would like to thank Sebastian Butschek and Laura Pohlan for helpful comments and suggestions. We further thank Maria Bidenko and Vanessa Lindenmaier who provided excellent research assistance.

1 Introduction

The use of administrative data sets in economics and especially labour economics has become more and more important for policy evaluation and empirical research. Apart from their large sample size, such data sets have the advantage of covering long time periods and offering precise longitudinal information on key variables such as earnings and labour market states. A further strength of most administrative data sets is that they mitigate problems of panel attrition that typically arise with survey data. In this study, we focus on the *BASiD* (2007) data set, which combines information from the German Pension Register and the Federal Employment Agency. The data provide longitudinal information on individuals' pension-relevant biographies up to the year 2007. Compared to other administrative data such as the *Integrated Employment Biographies* (*IEB*) from the German Federal Employment Agency, *BASiD* encompasses entire individual employment biographies (in general starting with the age of 14).

An important feature of the *BASiD* data is that it is particularly attractive for studying the Eastern German labour market. While recent studies based on administrative data have been restricted to the period after unification (see e.g. Kohn and Antonczyk, 2013), much of the literature dealing with the labour market in the German Democratic Republic (GDR) has relied on the German Socioeconomic Panel (*GSOEP*) (see e.g. Bird et al., 1994). These survey data provide retrospective GDR information for the years 1988 and 1989. A great advantage of *BASiD* over the *GSOEP* is that it contains full employment biographies of former GDR citizens prior to German unification. However, a shortcoming of *BASiD* is that it fails to provide information on individual covariates before 1992 (exceptions are gender and age), including the entire time period prior to unification. Given that especially information on educational attainment is of major relevance to many labour market applications, this study proposes different procedures for imputing the pre-unification education variable in the *BASiD* data.

While the precision of key variables such as earnings is generally considered to be high, educational information from administrative records is often subject to measurement error. In proposing an imputation procedure for the *BASiD* data, our paper is therefore related to the literature dealing with measurement error and missing values (Little and Rubin, 2014, Schafer, 2010, Manzari, 2004). While much of the literature on German administrative data has proposed procedures to eliminate inconsistencies and to impute missing values of earnings (Büttner and Rässler, 2008) or education (Fitzenberger et al., 2006, Wichert and Wilke, 2012), our approach has

to deal with a complete lack of educational information prior to unification. Instead of eliminating inconsistencies of an existing variable, we therefore have to provide a rule that indirectly infers educational information from other variables in the data set. We do so by exploiting information on education related periods that are creditable for the pension insurance. Combining these periods with information on the educational system in the former GDR, we propose three different imputation procedures. The first one, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The second one, *IMP2*, gives up the age constraints and aims to match somewhat broader education categories, into which the six categories in the *IEB* have been typically summarised in many empirical applications. Finally, the third one, *IMP3*, defines the education level based on potential years of education.

We validate our procedures by using external GDR census data documented by Steiner (1986) and Maaz (2002). These data allow us to compare the fraction of individuals in specific education-age group cells resulting from our imputations with the corresponding fractions from the census data for comparable cohorts in 1984. The general picture that emerges is that *IMP2* and *IMP3* tend to overpredict the share of medium-skilled workers and underpredict the share of high-skilled for all age groups, whereas they underpredict (overpredict) the share of low-skilled for the younger (older) age groups. The underprediction (overprediction) of the share of low-skilled for younger (older) age groups is also true for *IMP1*. Compared with *IMP2* and *IMP3*, *IMP1* gives rise to smaller deviations for the share of medium-skilled, whereas it overpredicts the share of high-skilled especially for the younger age groups. Overall, in terms of the trade-off between goodness of fit and data coverage, the more narrowly defined procedure *IMP1* performs best. However, once one re-classifies those with a GDR *Fachschule* degree as medium skilled workers, *IMP2* is to be preferred over *IMP1* and *IMP3*. After discussing potential sources of misclassification, we proceed by comparing our education variable prior to unification with educational information from the *IEB* right after unification. To do so, we first improve the *IEB* education information according to the algorithm proposed by Fitzenberger et al. (2006). For comparison purposes, we then follow Wichert and Wilke (2012) by performing a regression analysis in order to identify the importance of observables for a deviation of our imputed educational information from that in the *IEB*. Although regressions show that misclassification with respect to *IEB* information is clearly related to observables, we do not find any systematic pattern across skill groups.

The remainder of the paper is structured as follows. Section 2 describes the *BASiD* data. Section 3 describes the educational system of the GDR and discusses the three imputation rules. Section 4 validates the imputation results using external census data. Section 5 compares the results from our imputations with educational information from the *IEB* right after unification. The final Section 6 concludes.

2 *BASiD* Data

The *BASiD* data combine information from the *German Pension Register* with various data sources from the German Federal Employment Agency. In this contribution we use the scientific use file (*BASiD*-SUF) provided by the German Pension Insurance, which is a stratified random 0.25% sample.[1] This sample comprises all birth cohorts from 1940 to 1977[2], who have at least one entry in their social security records, leading to an overall sample of about 60,000 individuals. The sample has been drawn in a disproportionate manner and can be made representative using a weighting factor that is part of the data set (for a detailed description see Bönke, 2009, Himmelreicher and Stegemann, 2008). To identify former GDR citizens, we first select all individuals who prior to monetary union (i.e., the 30th of June 1990) had exhibited pension-relevant activities in the GDR, i.e. those individuals who had gained East German credit points prior to unification. This gives rise to a sample of 11,331 individuals with a total of 2,790,600 monthly spells.

The data provide longitudinal information on individuals' entire pension-relevant biographies up to the year 2007. Individual work histories cover the period from the year individuals were aged 14 until the age of 67. In Germany, statutory pension insurance is mandatory for all employees in the private and public sector, thus only excluding civil servants and self-employed individuals. In addition, contributions to the pension insurance are paid by the unemployment or health insurance during periods of unemployment and prolonged illness.

[1] A larger 1%-sample of the pension part of the data is available via on-site access at the Research Data Centre of the German Pension Insurance. An alternative version of the full *BASiD* 1%-sample (including also the data from the Federal Employment Agency) is also available at the Research Data Centre of the Institute for Employment Research (IAB). While the Pension Insurance version provides monthly spell information, the IAB version provides daily information for all pension relevant episodes (see Hochfellner et al., 2012).

[2] The cohort structure of our data implies that the earliest period in which we observe insured individuals is the year 1954, when those born in 1940 were 14 years old. During the subsequent years younger cohorts successively enter the data set, which gives rise to an increasingly mixed age structure. To ensure representativeness of the sample for the working-age population's age structure, we have constructed weights based upon administrative population data from the German Federal Statistical Office.

As stressed at the outset, the *BASiD* data provide an ideal basis for analysing labour market related questions of former GDR citizens for several reasons: First, it is the only German administrative data source that encompasses full employment biographies. In particular, the *Pension Register* contains information on all periods for which contributions were paid (employment, apprenticeship training, long-term illness, unemployment) as well as periods without contributions, but which were still creditable for the pension insurance. The latter refers to activities for which an individual receives pension credits, such as periods of school or university attendance after the age of 16 and periods of child rearing and caring.

Second, the *BASiD* data is the only individual level data set that contains employment biographies of former GDR citizens before German unification. After unification, former GDR citizens became entitled to transfer their pension-relevant activities to the FRG pension insurance system. For this purpose, the FRG Pension Insurance recorded all periods prior to unification which were creditable for the pension insurance (see above) as well as earnings up to the GDR social security cap. The pension data therefore allow researchers to track former GDR workers' entire pre- and post-unification employment histories up to the year 2007. Apart from the individual information on pension-relevant activities, the *Pension Register* provides information on age and gender.

Starting from 1975 in Western and from 1992 in Eastern Germany, employment spells subject to social security contributions from the *Pension Register* can be merged with data from the German Federal Employment Agency, the *Integrated Employment Biographies* (*IEB*) and the *Establishment History Panel*. The *Establishment History Panel* contains information on the establishment's workforce composition, establishment size as well as sector affiliation (see Table A.1 in the Appendix). Finally, the *IEB* provide further time-varying individual information on blue- or white-collar status, occupational status, educational status (see Table A.2 in the Appendix) and an establishment identifier. A shortcoming of the *BASiD* data is therefore that apart from age and gender most covariates are available after unification only, starting with the year 1992.

3 Education System and Imputation Procedures

3.1 The Education System in the GDR

Before we turn to our imputation procedures, we provide some institutional background information on the educational system in the former GDR. Due to their

common roots the educational systems in the former GDR and Western Germany (FRG) exhibit some similarities. However, in the course of the GDR's history, the systems considerably diverged with a strong effect on educational outcomes. While in Western Germany a first selection generally took (and still takes) place during primary school after four years of schooling, GDR pupils used to spend their first eight years together. Thus, completion of 8th grade in Eastern Germany may be considered the first official school degree as compared with 9th grade in Western Germany. The completion of 10th grade (referred to as "Polytechnische Oberschule" (*POS*)) was the second highest school-leaving degree in the former GDR. As will be discussed below, in the 1950s and 1960s only a minority obtained a *POS* degree, whereas in later years the majority of pupils was expected to attend school for ten years (Fuchs–Schündeln and Masella, 2016). After completion of *POS*, attending high school two more years (referred to as "Erweiterte Oberschule" (*EOS*)) gave rise to a degree equivalent to the Western German "Abitur". However, as a result of an increasing influence of the "state-governed labour force allocation", access to high school became highly limited since the late 1960s and was not only determined by pupils' performance but also by their own as well as their parents' political orientation (Hegelheimer, 1973; Huinink and Solga, 1994).

Regarding the quantitative relevance, the *POS* degree was of minor importance in early GDR years. In the 1960s, the fraction of pupils leaving school with 8th grade or less was about 50%. This share decreased gradually, until eventually completion of *POS* became the most common qualification. At the end of the socialist regime about 80% finished 10th grade (Maaz, 2002). Restrictive access to high school resulted in a stable low fraction of pupils completing high school of slightly above 10% (Solga, 2002). This development diverged considerably from that in Western Germany, where in 1989 a much higher share had attained lower educational qualification (30%) and a high school degree (25%) (Maaz, 2002).

As to vocational training, the length of an apprenticeship period was strongly determined by prior educational attainment from schooling. In general, vocational training after completing 8th grade took about 3 to 3.5 years, whereas an apprenticeship period following *POS* lasted only two years (Hegelheimer, 1973). Moreover, there was also the possibility to combine a three year apprenticeship training with the completion of a high school degree, which enabled individuals to enter a university afterwards. Given the low share of pupils with a highest degree of 8th grade or less, in 1987 about 78% of apprenticeship periods exhibited a duration of two years, 11% of 2.5 years and 11% of three years (Fuchs–Schündeln and Masella, 2016).

As a further possibility of post-secondary education, individuals could enter a so called *Fachschule*. Admission to a *Fachschule* was either possible after completion of *POS* or, alternatively, after completion of an apprenticeship training. The second type gave rise to a kind of technical university degree (equivalent to that in Western Germany), whereas the first type was closer to a vocational degree (Biermann, 2013). The average length of *Fachschule* period took about three years, whereas the length of university periods was about four years, on average (Krueger and Pischke, 1995).

3.2 Imputation Procedures

In what follows, we derive imputation rules that infer educational information from education related periods that are creditable for the pension insurance. As spelled out earlier, the *Pension Register* records periods for which contributions were paid (employment, long-term illness, unemployment) as well periods without contributions, but which are still creditable for the pension insurance, e.g. in terms of waiting times. While vocational training periods generally include periods with paid contributions, school and university episodes belong to the second type. By law, these latter periods are creditable for the pension insurance for individuals in full time education after the age of 16 for up to 8 years.

Combining the length of these periods with information on the educational system in the former GDR, we propose three different imputation procedures. The first one, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The six categories include (*ND*) no degree, (*HS*) a high school degree, (*VT*) a completed vocational training, (*VTHS*) a completed vocational training plus high school degree, (*TUD*) a technical university degree and finally, (*UD*) a university degree. The second one, *IMP2*, gives up the age constraints and aims to match the broader education categories, into which the six categories in the *IEB* are typically being summarised in many empirical applications: According to these, (1) low-skilled workers are those without any postsecondary degree, (2) medium-skilled workers have a completed apprenticeship training and (3) high-skilled workers obtained a degree from university or a technical university. Our last procedure, *IMP3*, simply defines these three categories based on potential years of education.

In general, the three different procedures involve a trade-off between precision and data coverage. *IMP1* bears the potential of losing information on individuals who exhibit school or vocational training spells, but not at the required age. By

giving up or loosening the age constraints, *IMP2* and *IMP3* may overcome this loss of observations at the expense of less precision. Table 1 summarises the constraints that *IMP1* imposes (for a graphical summary of the procedures see also Figure A.1. in the Appendix). For this procedure, we need to know at what ages the different degrees were typically completed in the socialist system. As documented by Krueger and Pischke (1995), 8th grade and 10th grade were passed at age 14 and 16, respectively. High school was generally completed at age 18.

According to *IMP1*, individuals are assigned "No degree" if they (1) exhibit a first employment spell and are younger than 17 years and/or (2) if they never experienced a school or apprenticeship spell throughout their observed history. Note that in case (1) the skill status may change once individuals have experienced a school or apprenticeship spell according to the rules spelled out below. Individuals having experienced a school episode between 1 and 1.5 years at the age of 17 or 18 are assigned a high school degree (category (*HS*) from the *IEB*). To match categories (*VT*) and (*VTHS*) from the *IEB*, we distinguish between those with a completed vocational training after 8th or 10th grade and those combining high school with apprenticeship training. Individuals are assigned to category (*VT*) (8th grade or 10th grade with apprenticeship) if they have experienced a vocational training spell at the age of 18 or younger with the length of the training lasting between 1.5 and 3.5 years.[3] As already mentioned, a further possibility was to combine a three year apprenticeship training with the completion of a high school degree. Thus, individuals having experienced an apprenticeship spell lasting between 2.5 to 3.5 years at the age of 19 or 20 are assigned an apprenticeship plus high school degree (*VTHS*).

In the *IEB* data, the high-skilled comprise those with a technical university degree and a university degree. In what follows, we will simply match the *IEB* technical university degree with a GDR technical school (*Fachschule*) degree.[4] Our assignment rule relies on the fact that completion of a *Fachschule* took three years, whereas the completion of a university degree took four years on average. To capture the completion of a *Fachschule* degree after *POS*, individuals having experienced a schooling episode of at least 2 up to 3.5 years at the age of 19 to 20 are assigned (*TUD*). The same assignment is made for individuals having experienced a schooling episode of at least 1.5 up to 3.5 years at the age of 20 or older.[5] As we observe a

[3]Note that we also assign individuals younger than 17 to category (*VT*) if they have experienced apprenticeship spells of less than three years to account for the possibility of potential underreporting of vocational training spells in the data set.

[4]We will discuss the limitations of such an approach below.

[5]Note that for this group we also count schooling spells of less than two years. The reason is

Table 1: Imputation Procedure (1)

Category	Characteristics	Criteria	# Individuals
No Degree	Age	≤ 17	1,336
	First socio-economic spell	employed	
	or		
	School/Apprenticeship spell	none	
High School (Abitur)	Age	17-18	274
	&		
	School spell	1.0 - 1.5 years	
Vocational Training	Age	≤ 18	3,763
	&		
	Apprenticeship spell	1.5-3.5 years	
Vocational Training and High School	Age	19-20	644
	&		
	Apprenticeship spell	2.5 - 3.5 years	
Technical University Degree	Age	19-20	1,421
	&		
	School spell	2 - 3.5 years	
	Age	≥ 20	
	&		
	School spell	1.5 - 3.5 years	
	Age	≥ 20	
	&		
	Apprenticeship spell	> 2.5 years	
University Degree	Age	≥ 22	940
	&		
	School spell	≥ 3.5 years	

Source: BASiD 2007.
Notes: The total number of individuals under consideration (before 1992) is 11,331. The coverage rate (covered individuals = 7,543/total individuals = 11,331) of Imputation Procedure (1) is 67%. Note that the number of covered individuals is less than the sum over all categories because of transitions to a higher educational level. We allow for interruptions of the school and apprenticeship spells and for continuing the spell's duration after the interruption. There are 2,191 (19%) individuals in the sample with interruptions in school or apprenticeship spells with a average length of 11 months. About 50% of all interruptions occur within the first 3 months. In 50% of all cases an interruption occurs because of sickness, 26% of regular employment and 15% of child care. See also for a graphical illustration of the procedures and a detailed description of interruptions Appendix A, Figure A.1. After applying the criteria we extrapolate the educational degree to future spells until we observe a change in individuals' educational status.

considerable fraction of individuals having experienced an apprenticeship period of at least 2.5 years after the age of 20, those individuals are assigned a *Fachschule* degree as well. This is to account for the possibility that a *Fachschule* degree, which was frequently associated with the completion of a vocational training, might have been

that - given the predetermined educational biographies - any experience of a schooling spell at the age of 20 or older is likely to reflect a higher educational degree.

misreported as a vocational training period in the Pension data. Finally, individuals are assigned a university degree (*UD*) if they have experienced a schooling spell of at least 3.5 years at the age of 22 years or older.

Procedure (2) is summarised in Table 2. Low skilled workers are those assigned "No degree" (see *IMP1*). Medium skilled-workers need to have at least 1.5 years of formal apprenticeship training, whereas high skilled workers are those with school spells of at least three years.

Table 2: Imputation Procedure (2)

Category	Characteristics	Criteria	# Individuals
Low-skilled	Age	≤ 17	1,336
	First socio-economic spell	employed	
	or		
	school/apprenticeship spell	none	
Medium-skilled	Apprenticeship (cumulative)	> 1.5 years	7,889
High-skilled	School (cumulative)	> 3 years	1,367

Source: BASiD 2007.
Notes: The total number of individuals under consideration (before 1992) is 11,331. The coverage rate (covered individuals = 10,082/total individuals = 11,331) of Imputation Procedure (2) is 91%. Note that the number of covered individuals is less than the sum over all categories because of transitions to a higher educational level. Cumulative spells allow for interruptions and for continuing the spell's duration after the interruption. After applying the criteria we extrapolate the educational degree to future spells until we observe a change in individuals' educational status.

The third approach relies on potential years of education (*IMP3*). Given that formal unemployment was officially barely present in the GDR[6], the idea of this rule is that the first employment spell should have immediately followed the completion of an educational degree. We define individuals as low-skilled if they are less than 17 years old and are labeled as employed (employment subject to social security contributions excluding apprenticeship periods). Individuals starting employment between age 17 and 20 are defined as medium-skilled, whereas high-skilled individuals are those with a first employment spell at the age of 21 to 28.[7] Note that this approach does not account for potential changes in educational attainment over individuals' life courses.

[6]See a discussion by Gürtler et al. (1990) on hidden unemployment.

[7]In total the procedure generates 1,535 low-skilled individuals, 9,006 medium-skilled and 586 high-skilled. The *IMP3* approach has a coverage rate of 99%.

4 Qualification Structure

In what follows, we attempt to externally validate our proposed imputation procedures. To do so, we compare the qualification structures by age groups obtained from our procedures with those from census data from 1981 for the whole residential population (Steiner, 1986 and Maaz, 2002). The first three panels in Table 3 show the imputed qualification structures distinguished by five age groups, whereas the last panel displays the figures from the census data. Note that our imputed figures refer to the same age groups three years later in 1984 as the cohort structure of our data set only allows us to provide figures for the oldest age group (40-44) from 1984 onwards. The census figures shown in the bottom panel indicate that the share of low-skilled workers is U-shaped, whereas the share of medium-skilled workers is decreasing with age. For high-skilled individuals, we observe a kind of inverse U-shaped picture with the largest share of high-skilled individuals in the second oldest age group. For the younger age groups, *IMP1* assigns substantially more individuals to the high-skilled group compared to the census figures and less to the medium and low-skilled category. For the older age groups, it assigns substantially more individuals to the low-skilled and less to the medium and high-skilled category. *IMP2* generates high and low-skilled (medium-skilled) shares that are substantially lower (higher) for most of the age groups than those from the census data. *IMP3* leads to an even more pronounced underprediction of the share of high-skilled in all age groups. This reflects that *IMP3* does not account for completed school episodes at the age of 21 or younger that might have led to a *Fachschule* (technical school degree) after completion of *POS*. Looking at the low-skilled share obtained from *IMP3* suggests that the average age of the first labor market spell increased over time. This is due to the fact that the dominant school degree moved from 8th grade or less (about 30% of school-leavers had less than 8th grade in the 50s and up to 60% had 8th grade) to the 10th grade over that time period (for a detailed description see Solga, 2002).

Overall, the discrepancies are non-negligible for each of our proposed procedures. To rank the procedures in terms of the implied deviations from the census data, we calculate the sum of squared differences between the imputed and the census figures over the different age-skill cells. Table 4 presents the results. *IMP1* results in a sum of squared differences equal to 1105, whereas *IMP2* involves a sum of squared differences equal to 2189. Using *IMP3* we obtain a number of 3791. Based on these numbers, we would prefer the first procedure over the second and the last one. However, the different procedures capture different numbers of observations.

Table 3: Qualification structure by age groups - 1984, part 1

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20-24	13.9	62.0	24.1
	25-29	10.5	55.8	33.7
	30-34	14.6	60.1	25.3
	35-39	25.3	52.5	22.3
	40-44	31.2	52.3	16.5
IMP2	20-24	5.6	89.5	4.9
	25-29	5.9	78.7	15.4
	30-34	9.9	75.7	14.4
	35-39	16.9	73.9	9.1
	40-44	29.4	63.5	7.1
IMP3	20-24	7.0	91.5	1.5
	25-29	7.1	86.5	6.4
	30-34	12.1	81.2	6.7
	35-39	16.8	78.7	4.6
	40-44	34.2	62.8	3.0
Census Data	20-24	18.0	71.0	11.0
	25-29	12.0	65.0	23.0
	30-34	11.0	63.0	26.0
	35-39	11.0	60.0	29.0
	40-44	17.0	60.0	23.0

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are documented by Steiner (1986) and Maaz (2002). Low-skilled individuals in the census data are individuals without any degree or with a partial completion of a vocational training. Medium-skilled workers include those with a completed vocational training and so-called "Meister", whereas high-skilled workers consist of those with a technical school degree and a university degree. It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. We perform statistical significance test (t-test) between the generated variables for each age group and the census data. According to *IMP1*, only the share of high-skilled among those 30-34 years old does not significantly differ from the census data. For *IMP2* the share of low-skilled in the age group 30-34 years and the share of medium-skilled in the age group 40-44 years do not significantly differ from the census data. *IMP3* shows that only the share of low-skilled in the age group 30-34 years does not significantly differ from the census data.

Table 4: Squared Differences among Imputation Procedures - 1984, part 1

	Sum of squared Differences Q	Individuals N	Coverage C	Q/N	Q/C
IMP1	1105	6,398	0.60	0.173	1842
IMP2	2189	8,323	0.78	0.263	2806
IMP3	3791	9,386	0.88	0.404	4308

Source: BASiD 2007, weighted statistics.
Notes: Total number of individuals in 1984: 10,702. The coverage rate of *IMP3* is rather low because for some individuals the requirements were not met at that point in time.

Dividing the sum of squared differences by the number of individuals (or equivalently by the coverage rate), *IMP1* is still preferred over *IMP2* and *IMP3*.[8]

What still remains to be resolved is the question as to why we observe these differences, especially for the share of high-skilled. One possible explanation for the strong deviations produced by *IMP2* and *IMP3* could be that the census data are biased towards an over-reporting of high-skilled individuals. An alternative explanation could be that there is a reporting error for school and apprenticeship spells in the administrative data set. Regarding the second explanation (reporting error) we do not think that this is a major concern as especially *IMP2* and *IMP3* should also capture those, whose pension accounts in terms of schooling and apprenticeship episodes might have been incomplete.

Moreover, note that even procedure *IMP1*, which explicitly attempts to distinguish those with a *Fachschule* from those with a university degree, substantially underpredicts (overpredicts) the share of high-skilled especially for the older (younger) age groups. To obtain a more precise picture of potential misclassification sources, we break down the results from *IMP1* by female and male workers as well as by those with a technical school and university degree (see Table 5). The resulting figures show that the way we assigned the education variable performs relatively well for university graduates (for both male and female workers) and medium-skilled male workers as compared to the technical school degree, where large discrepancies can be observed for all age groups. To account for a potential misclassification of a technical school as a vocational degree, one approach to handle this difficulty could be to re-classify the medium-skilled by assigning all technical school graduates to the medium-skilled. After this re-classification, the high-skilled group would only include university graduates. In terms of the Western German skill categories, such a re-classification could e.g. be justified by the fact that in the GDR individuals could enter a technical school after completion of *POS* (Krueger and Pischke, 1995), which would be rather equivalent to a Western German vocational training degree. The results from this re-classification are shown in Table 6. The upper part of Table 6 shows that the share of high-skilled workers decreases and becomes closer to the official data particularly for the older age groups. Thus, treating former *Technical School* graduates as medium-skilled may help to draw a somewhat clearer picture for the high-skilled group. After the re-classification, the sum of squared differences becomes smaller for all three imputation procedures with *IMP2* and *IMP3* showing

[8]Instead of using the sum of all squared differences, we perform the same exercise using the absolute difference. The ranking remains unchanged.

Table 5: Qualification structure by age groups and gender - 1984

	Age	Medium-skilled		Technical School		University	
	group	male	female	male	female	male	female
IMP1	20-24	72.0	54.9	15.6	29.4	0.0	1.6
	25-29	62.5	52.0	17.1	18.7	6.6	10.0
	30-34	65.1	55.3	12.1	10.8	10.3	8.0
	35-39	58.7	47.9	11.7	14.6	10.0	3.5
	40-44	63.9	46.7	9.0	9.6	9.3	2.9
Census data	20-24	78.0	64.0	1.0	16.0	1.0	3.0
	25-29	72.0	59.0	6.0	19.0	10.0	10.0
	30-34	67.0	58.0	11.0	20.0	13.0	9.0
	35-39	63.0	56.0	16.0	22.0	13.0	6.0
	40-44	62.0	57.0	15.0	16.0	12.0	5.0

Source: BASiD 2007, weighted statistics.
Notes: The census data cover the residential population in 1981 and are documented by Steiner (1986) and Maaz (2002). It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. The share "Technical School" always statistically differs from that in the census data. The fraction "University degree" statistically differs from that in the census data only for males in the age categories 20-24 and 25-29 and for females in the age categories 20-24 and 35-39.

Table 6: Qualification structure by age groups - 1984, part 2

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20-24	13.9	84.7	1.4
	25-29	10.5	75.5	14.0
	30-34	14.6	71.8	13.6
	35-39	25.3	66.0	8.7
	40-44	31.2	62.9	5.9
Census Data	20-24	18.0	80.0	2.0
	25-29	12.0	78.0	10.0
	30-34	11.0	79.0	10.0
	35-39	11.0	79.0	10.0
	40-44	17.0	75.0	8.0

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are documented by Steiner (1986) and Maaz (2002). It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. The medium-skilled group now includes former technical school graduates. A t-test for statistical differences with respect to the census data shows that the low- and medium-skilled shares in the second oldest age group as well as the share of high-skilled in the two oldest age groups are not statistically different from zero.

stronger improvements.[9] The ranking of the procedures change as well. Based on

[9]Note that for *IMP2* and *IMP3* the improvement only results from re-classifying the census data.

Table 7: Squared Differences among Imputation Procedures - 1984, part 2

	Sum of squared Differences Q	Individuals N	Coverage C	Q/N	Q/C
IMP1	869	7,297	0.64	0.119	1358
IMP2	699	9,681	0.85	0.072	822
IMP3	912	10,823	0.96	0.084	950

Source: BASiD 2007, weighted statistics.
Notes: Total number of individuals in 1989: 11331.

the sum of squared differences and the number of individuals (coverage rate), *IMP2* is now preferred over *IMP3* and *IMP1*.

Given the substantial deviations for the older age groups that result from all three procedures, we next check whether our procedures are at least able to reproduce the documented decline in the fraction of those leaving school at 8th grade or less. We do this by estimating our skill shares at a later point in time (five years later in 1989). Individuals from the oldest age group (those aged 44) in 1989 were born in 1945 and were potentially available for the labour market in 1960/1961. Given that the share of those leaving school at 8th grade or less was twice as high in the 1950s as compared to the 1960s (Solga, 2002), we expect a sharp decline in the predicted low-skilled share especially for older age groups. Table 8 presents the results for *IMP1*. As can be seen, our expectations are borne out by the figures, which are also

Table 8: Qualification structure by age groups - 1989

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20-24	9.1	88.6	2.3
	25-29	9.9	77.3	12.7
	30-34	10.7	76.3	13.1
	35-39	15.3	72.7	11.9
	40-44	25.8	65.9	8.3
Census Data	20-24	18.0	80.0	2.0
	25-29	12.0	78.0	10.0
	30-34	11.0	79.0	10.0
	35-39	11.0	79.0	10.0
	40-44	17.0	75.0	8.0

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are documented by Steiner (1986) and Maaz (2002). It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. The medium-skilled group now includes former technical school graduates. Statistical significance test are available upon request.

in line with the descriptive statistics shown in Solga (2002). (Unreported) results reveal that the share of low-skilled obtained from the other two procedures drop by a similar magnitude. Thus, our procedures at least appear to provide consistent results in terms of the decline of those leaving school at 8th grade or less.

Taken together, our comparison with the census data suggests that all three procedures exhibit non-negligible deviations from our external data source. In order to finalise our decision about which procedure to use, we further compare our imputation results to information provided by the *IEB* subpart of the data set.

5 Comparison with educational information from the *IEB*

We next compare the results from our imputation procedures with educational information from the *IEB*, which can be merged to the *BASiD* data. We use January 1992 as a reference point as *IEB* information is available from 1992 onwards for Eastern Germany. With this comparison we have to keep in mind that the *IEB* education information may be subject to measurement error as well. To mitigate this issue, we correct the *IEB* education information using the imputation algorithm *IP1* proposed by Fitzenberger et al. (2006). The authors suggest three different imputation rules without a strict order. The idea behind their rules is based on the assumption that individuals cannot lose their educational degrees. In what follows we use *IP1*, since according to Wichert and Wilke (2012) imputation procedure 1 (*IP1*) leads to a stronger reduction in measurement error.

We perform this exercise for all of our imputation procedures. Table 9 first cross tabulates the results from procedure *IMP1* using the categories described in Table 1 with education information from the *IEB*. Table 9 shows that the best match is obtained for the vocational training (*VT*) category with a fraction of 56% receiving this category from both our imputation procedure *IMP1* and the *IEB*. In contrast, among those assigned a vocational training plus high school degree (*VTHS*) in the Pension data, over 50% exhibit only a vocational training degree (*VT*) in the *IEB*. Note that this may either reflect that our imputation procedure wrongly assigns a vocational training plus high school degree or, alternatively, that the GDR high school degree has either not been reported or recognised by Western German employers.

Table 9: Cross tabulation of *IMP1* vs. *IEB*, part 1

	Degree *IMP1*					
IEB (*IP1*)	Missing	ND	VT	VTHS	TUD	UD
Missing	**37.3**	45.0	38.0	37.3	33.6	27.6
ND	5.5	**12.8**	4.2	3.4	1.8	0.4
VT	53.4	39.8	**55.6**	53.4	40.9	18.0
VTHS	1.1	1.0	0.8	**2.7**	6.0	5.9
TUD	1.8	0.9	1.3	1.8	**13.7**	9.3
UD	0.7	0.5	0.2	1.3	3.6	**38.4**
Observations	3788	1179	3575	560	1148	938

Source: BASiD 2007, weighted statistics.
Notes: The category *High School* plays with 1.2% a minor role and is not presented in the table. Total number of observations is 11,331. The reference point in time for the comparison is January 1992. Abbreviations: ND: no degree, VT: completed vocational training, VTHS: high school with vocational training. TUD: technical university degree, UD: university degree. In the Pension data *TUD* corresponds to a technical school (*Fachschule*) degree.

Moreover, the *IEB* comparison also produces large deviations for those assigned no degree in the Pension data (*ND*), who mostly exhibit a vocational training (*VT*) in the *IEB*. A potential explanation would be that our defined rules from procedure *IMP1* might have changed over time, such that older workers could have completed an apprenticeship within a shorter duration. If this was the case, the deviation of *IMP1* from *IEB* information should be mitigated using *IMP2*, as this procedure requires only a cumulative apprenticeship spell of at least 1.5 years. The second panel of Table 10 shows that this only accounts for a small part of the observed deviation in Table 9. In Table 10 the share of those assigned no degree (*ND*) in the Pension data, who exhibit a vocational training (*VT*) in the *IEB*, slightly decreases from 42% to 40% under *IMP2*. Note that the share of low-skilled decreased substantially for all age groups between 1984 and 1989, such that over-reporting a low-skilled status is unlikely to explain the deviation for the low-skilled in Table 9. An alternative explanation would be vocational on-the-job training, which generally cannot be ruled out as a source of any deviation between the (imputed) pension data and *IEB* information.

Given the problem with correctly assigning the technical school (or technical university degree (*TUD*)), over 45% of those assigned *TUD* in the Pension data exhibit only a vocational training degree (*VT* or *VTHS*) in the *IEB*. This highlights again the difficulty in distinguishing between a GDR technical school and a vocational degree. Note that this misclassification might also reflect the fact that educational degrees obtained during GDR times might have not been recognised by

Western German employers after unification. The overlap of *TUD* with a university degree (*UD*) is only moderate with about 4%.[10] From the 938 individuals who were assigned *UD* in the Pension data, about 38% exhibit also a *UD* in the *IEB*. 28% have missing values in the *IEB* and about 18% intersect with vocational training (*VT*).

The cross tabulation of the results from *IMP2* and *IMP3* with education information from the *IEB* is shown in the next table. As these procedures target three education categories - low-skilled, medium-skilled and high-skilled, we also provide the shares for the first imputation procedure *IMP1* using the three categories in the upper panel. The picture that emerges from Table 10 is that all three procedures

Table 10: Cross tabulation of *IMP1*, *IMP2* and *IMP3* vs. *IEB*, part 2

		IMP1		
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**37.3**	43.6	37.9	30.9
Low-skilled	5.6	**11.9**	4.1	1.5
Medium-skilled	54.5	42.4	**56.3**	36.6
High-skilled	2.6	2.2	1.7	**30.9**
Observations	3788	1322	4135	2086
		IMP2		
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**37.6**	45.3	37.3	28.5
Low-skilled	11.6	**12.9**	3.5	0.9
Medium-skilled	44.8	40.3	**57.0**	28.6
High-skilled	6.0	1.5	2.3	**42.1**
Observations	1249	1181	7539	1362
		IMP3		
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**23.5**	42.7	36.8	29.5
Low-skilled	54.4	**9.5**	3.8	0.9
Medium-skilled	22.1	43.3	**53.3**	31.6
High-skilled	0.0	4.4	6.1	**38.1**
Observations	136	1535	9074	586

Source: BASiD 2007.
Notes: Total number of observations 11,331. The reference point in time for the comparison is January 1992. Abbreviations (IEB): Low-skilled include ND and HS, Medium-skilled include VT and VTHS, High-skilled include TUD and UD.

show again very similar patterns. The best match is obtained for the medium-skilled, whereas the worst match results for the low-skilled.[11] Overall, for the 11,331

[10]This again suggests that re-classifying a *Fachschule* degree as medium-skilled would potentially reduce measurement error issues and misclassification.

[11]Note that misclassification also occurs when comparing different sources from the *IEB*.

individuals in the *BASiD* data set we obtain the highest number of missing values (3,788) for procedure *IMP1*. The number of missing values in the *IEB* (4,204) variable is, however, 15% larger compared to *IMP1* and twice as large compared to *IMP2*, indicating a substantial gain in information resulting from our imputation procedures. As mentioned earlier, there exists generally a trade-off between precision and coverage in terms of missing values. Given that the main diagonal values are largest for *IMP2*, which simultaneously reduces the number of missing values by almost 75%, procedure *IMP2* seems to provide a quite reasonable compromise between matching *IEB* information and data coverage.

To analyse whether any deviation from *IEB* information is systematically related to observables, we next perform a logit analysis using the results from procedure (*IMP2*) for the year 1992. The dependent variable is one if the educational degree assigned in the Pension data deviates from the educational information in the *IEB* and is zero otherwise. Table 11 presents the estimated marginal effects for the probability of misclassification in the sample separately by skill group. While the results suggest that misclassification is related to some observables, there appears to be no systematic pattern across skill groups. In particular, the marginal effects of the covariates vary considerably across skill groups. Being female increases the probability of being misclassified by about 18 p.p. for the high-skilled. Younger individuals have a higher probability of being misclassified if they are assigned low-skilled in the Pension data (Column (2)), whereas there is no significant relation for the other groups. Individuals employed by smaller establishments are more likely to be misclassified in all skill groups than in the reference group of medium-sized establishments. Individuals employed by large establishments exhibit a higher probability of misclassification only if they are assigned medium-skilled (Column (3)), whereas the marginal effects for low and high-skilled in the Pension data (Column (2) and (4)) are not significant. The signs of the marginal effects of different industry affiliations do not reveal any systematic pattern either and also vary greatly across skill groups.

6 Conclusions

The *BASiD* data set provides the only available data source that contains full employment biographies of former GDR citizens prior to German unification. How-

Wichert und Wilke (2012) compare job seekers' histories (BewA) with data from the Employment Register (BeH) and obtain high misclassified results (Wichert and Wilke, 2012). For example, the match between BewA and BeH data for the technical school degree variable is about 36% after correcting the variable by using the imputation algorithm.

Table 11: Marginal effects of a logit regression

Dependent variable:	Deviation from *IEB* information of ...			
	(1)	(2)	(3)	(4)
	Missing	Low-skilled	Medium-skilled	High-skilled
female	-0.032 (0.020)	0.032 (0.039)	0.022 (0.014)	**0.177** (0.035)
Age group				
< 25 (ref.)				
25 - 30	-0.017 (0.042)	**0.115** (0.050)	-0.033 (0.022)	0.028 (0.075)
30 - 35	0.005 (0.037)	-0.045 (0.096)	0.020 (0.024)	0.077 (0.071)
35 - 40	-0.032 (0.041)	0.015 (0.052)	0.016 (0.024)	-0.036 (0.072)
40 - 45	-0.089 (0.052)	-0.054 (0.084)	-0.002 (0.023)	0.012 (0.074)
45 - 50	-0.038 (0.045)	-0.058 (0.041)	-0.005 (0.024)	0.024 (0.079)
Firm size				
Medium size (ref.)				
below 20	-0.035 (0.053)	**0.128** (0.035)	**0.146** (0.032)	**0.279** (0.062)
20 - 49	-0.020 (0.062)	-0.042 (0.064)	0.033 (0.036)	-0.086 (0.070)
200 - 999	-0.002 (0.041)	-0.047 (0.044)	0.006 (0.027)	0.000 (0.055)
1000 and above	**-0.110** (0.051)	0.059 (0.041)	**0.287** (0.026)	0.099 (0.056)
Economic sector				
Construction (ref.)				
Agrar	**0.022** (0.006)	-0.001 (0.009)	**-0.011** (0.005)	0.003 (0.012)
Energy/Mining	**-0.166** (0.072)	0.072 (0.046)	**0.258** (0.030)	**0.203** (0.086)
Manufacturing	**0.067** (0.021)	-0.042 (0.049)	**-0.089** (0.018)	**-0.234** (0.063)
Wholesale	-0.051 (0.077)	**0.123** (0.039)	**-0.071** (0.026)	0.087 (0.106)
Traffic/communic.	-0.049 (0.077)	**0.095** (0.041)	**-0.105** (0.026)	0.050 (0.129)
Banking/insurance	-0.109 (0.134)	-	-0.008 (0.072)	-0.068 (0.126)
Other services	-0.026 (0.043)	-0.002 (0.057)	**0.086** (0.033)	**-0.133** (0.062)
Non-profit	-0.067 (0.117)	-	**-0.109** (0.038)	-0.021 (0.114)
Public sector	-0.018 (0.059)	0.022 (0.065)	-0.016 (0.037)	0.070 (0.075)
Predicted prob.	0.917	0.843	0.239	0.465
Log. likelihood	-260.0	-384.6	-2109.9	-567.8
Observations	734	587	4394	910

Source: BASiD 2007.
Notes: Robust standard errors are in parentheses. Bold numbers represent significance level of 5%.

ever, a shortcoming of *BASiD* is that it fails to provide information on individual covariates prior to unification (exceptions are gender and age). Given that especially information on educational attainment is of major relevance to many labour market applications, this study proposes different procedures for imputing the pre-unification education variable in the *BASiD* data.

Our proposed procedures exploit information on education related periods that

are creditable for the pension insurance. Combining these periods with information on the educational system in the former GDR, we investigate three different imputation procedures. The first one, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The second one, *IMP2*, gives up the age constraints and aims to match somewhat broader education categories, into which the six categories in the *IEB* have been typically summarised in many empirical applications. Finally, the third one, *IMP3*, defines the education level based on potential years of education.

We validate our procedures by using external GDR census data documented by Steiner (1986) and Maaz (2002). These data allow us to compare the fraction of individuals in specific education-age group cells resulting from our imputations with those from the census data for comparable cohorts in 1984. The general picture that emerges is that *IMP2* and *IMP3* tend to overpredict the share of medium-skilled workers and underpredict the share of high-skilled for all age groups, whereas they underpredict (overpredict) the share of low-skilled for the younger (older) age groups. The latter is also true for *IMP1*. Compared with *IMP2* and *IMP3*, *IMP1* gives rise to smaller deviations for the share of medium-skilled, whereas it overpredicts the share of high-skilled especially for the younger age groups. Overall, when balancing out the trade-off between goodness of fit and data coverage, the more narrowly defined procedure *IMP1* performs best.

Finally, a comparison of our (imputed) education information prior to unification with educational information from the *IEB* right after unification suggests that the best fit is obtained for those assigned a vocational training degree in the Pension data. The largest discrepancy is observed for those assigned a technical university degree in the Pension data of whom a large fraction (over 50%) exhibits only a vocational training degree in the *IEB*. This highlights again the difficulty in distinguishing between a GDR *Fachschule* and a vocational training degree. A simple approach to handle this difficulty could be to redefine the medium-skill category by assigning all technical (school) university graduates to the medium-skilled, such that the high-skilled group would only include university graduates. After doing so, *IMP2* would be preferred over *IMP1* and *IMP3*. Such a re-classification could reduce measurement error issues and misclassification, as long as it would be performed consistently over the whole sample period.

References

Biermann, H. (2013), *Berufsausbildung in der DDR: zwischen Ausbildung und Auslese*, Springer-Verlag.

Bird, E. J., Schwarze, J. and Wagner, G. G. (1994), 'Wage Effects of the Move toward Free Markets in East Germany', *Industrial and Labor Relations Review* **47**(3), 390–400.

Bönke, T. (2009), 'Gekappte Einkommen in prozessgenerieten Daten der Deutschen Rentenversicherung: Ein paretobasierter Imputationsansatz', *Deutsche Rentenversicherung Bund (Eds.): DRV-Schriften* **55**(2009), 214–230.

Büttner, T. and Rässler, S. (2008), 'Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity', *IAB Discussion Paper, No. 2008, 44* .

Fitzenberger, B., Osikominu, A. and Völter, R. (2006), 'Imputation rules to improve the education variable in the IAB employment subsample', *Schmollers Jahrbuch: Journal of Applied Social Science Studies* **126**(3), 405–436.

Fuchs–Schündeln, N. and Masella, P. (2016), 'Long-Lasting Effects of Socialist Education', *Review of Economics and Statistics (forthcoming)* .

Gürtler, J., Vogler Ludwig, K. and Ruppert, W. (1990), *Verdeckte Arbeitslosigkeit in der DDR*, Ifo-Institut für Wirtschaftsforschung.

Hegelheimer, A. (1973), 'Berufsausbildung in der DDR', *Gewerkschaftliche Monatshefte* **24**(3), 179–193.

Himmelreicher, R. and Stegemann, M. (2008), 'New possibilities for socio-economic research through longitudinal data from the Research Data Centre of the German Federal Pension Insurance', *Journal of Applied Social Science Studies* **128**(4), 647–660.

Hochfellner, D., Müller, D. and Wurdack, A. (2012), 'Biographical data of social insurance agencies in Germany–improving the content of administrative data', *Schmollers Jahrbuch* **132**, 443–451.

Huinink, J. and Solga, H. (1994), 'Occupational Opportunities in the GDR: A Privilege of the Older Generations?', *Zeitschrift für Soziologie* **23**(3), 237–253.

Kohn, K. and Antonczyk, D. (2013), 'The aftermath of reunification', *Economics of Transition* **21**(1), 73–110.

Krueger, A. B. and Pischke, J.-S. (1995), 'A Comparative Analysis of East and West German Labor Markets: Before and After Unification', *in: Freeman und Katz: Differences and Changes in Wage Structures,* pp. 405–446.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Maaz, K. (2002), 'Ohne Ausbildungsabschluss in der BRD und DDR: Berufszugang und die erste Phase der Erwerbsbiographie von Ungelernten in den 1980er Jahren', *Working Paper of the Independent Research Group of Max-Planck-Institute for Educational Research No. 3/2002* .

Manzari, A. (2004), 'Combining editing and imputation methods: an experimental application on population census data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **167**(2), 295–307.

Schafer, J. L. (2010), *Analysis of incomplete multivariate data*, CRC press.

Solga, H. (2002), 'Jugendliche ohne Schulabschluss und ihre Erwerbsbiografien', *in: Schnabel et al. (Hrsg.): Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick* .

Steiner, I. (1986), 'Struktur der Allgemeinausbildung und Berufsausbildung der Wohnbevölkerung in der DDR - Berufs- und Bildungsweglaufbahnen von Schulabsolventen', *Akademie der Pädagogischen Wissenschaft der DDR* .

Wichert, L. and Wilke, R. A. (2012), 'Which factors safeguard employment?: an analysis with misclassified German register data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175**(1), 135–151.

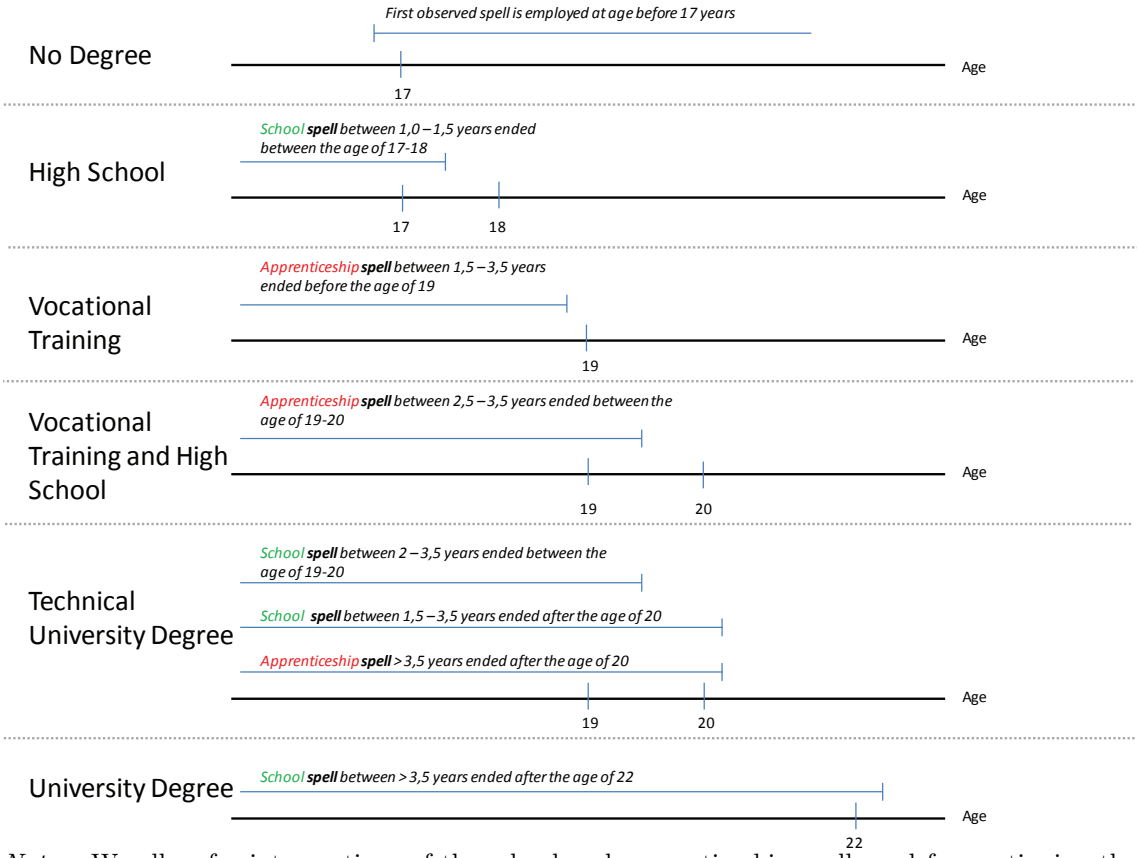
Appendix: Data Description

Variable	Definition/Categories:
Establishment size	Size $<$ *20*
	20 $\leq$ Size $<$ *50*
	50 $\leq$ Size $<$ *200*
	200 $\leq$ Size $<$ *1000*
	Size $\geq$ *1000*
Workforce composition	Share of employees younger than 30 years
	Share of employees older than 50 years
	Share of low-skilled employees
	Share of female employees
Sector affiliation	Agriculture/Forestry
	Mining and manufacturing
	Energy/Water supplies
	Construction
	Wholesale and retail trade
	Transport and communication
	Financial intermediation
	Other service activities
	Public administration

Table A.1: Definition of establishment characteristics gained from the *Employment Statistics Register*

Variable/Categories	Definition
GDR-Spell	GDR spells are identified based on the regional origin (*Beitrittsgebiet*) of the pension contributions
Educational Status 6 Categories	
NO DEGREE (*ND*)	No secondary level degree
HIGH SCHOOL	High school degree (Abitur)
VOC. TRAINING (*VT*)	Completed vocational training
VOC. TRAINING + HIGH SCHOOL (*VTHS*)	Completed vocational training plus high school
TECH. UNIVERSITY (*TUD*)	Fachschule or Technical University Degree
UNIVERSITY (*UD*)	University Degree
Educational Status 3 Categories	
LOW-SKILLED	No degree or high school degree
MEDIUM-SKILLED	Completed vocational training
HIGH-SKILLED	Technical college degree or university degree

Table A.2: Definition of individuals characteristics



Notes: We allow for interruptions of the school and apprenticeship spells and for continuing the spell's duration after the interruption. The distribution of interruptions differ by school or apprenticeship episodes. Regarding apprenticeships episodes, there are 1,252 individuals in the sample with interruptions in apprenticeship spells with a average length of 4 months. About 75% of all interruptions occur within the first 3 months. In 77% of all cases an interruption occurs because of sickness, 10% of regular employment and 11% of child care. The average age at the start of the interruption is 17.6 years. Regarding school episodes, there are 954 individuals in the sample with interruptions in school spells with a average length of 19 months. About 50% of all interruptions occur within the first 12 months. In 13% of all cases an interruption occurs because of sickness, 50% of regular employment and 20% of child care. The average age at the start of the interruption is 20 years. Given the longer durations of school spell interruptions and the average age at the start of the interruptions of 20 years, it is most relevant for the technical university degree and university degree as the end of a school episode is defined to be after 20 years and 22 years, respectively.

Figure A.1: Graphical Illustration of *IMP1*