Hutter, Christian; Möller, Joachim; Penninger, Marion

Article

Reducing the Need for Heuristic Rules – An Iterative Algorithm for Imputing the Education Variable in SIAB

Schmollers Jahrbuch – Journal of Applied Social Science Studies. Zeitschrift für Wirtschafts- und Sozialwissenschaften

Provided in Cooperation with:
Duncker & Humblot, Berlin

Suggested Citation: Hutter, Christian; Möller, Joachim; Penninger, Marion (2015) : Reducing the Need for Heuristic Rules – An Iterative Algorithm for Imputing the Education Variable in SIAB, Schmollers Jahrbuch – Journal of Applied Social Science Studies. Zeitschrift für Wirtschafts- und Sozialwissenschaften, ISSN 1865-5742, Duncker & Humblot, Berlin, Vol. 135, Iss. 3, pp. 355-388, https://doi.org/10.3790/schm.135.3.355

Schmollers Jahrbuch 135 (2015), 355 – 388
Duncker & Humblot, Berlin

Reducing the Need for Heuristic Rules – An Iterative Algorithm for Imputing the Education Variable in SIAB

By Christian Hutter, Joachim Möller and Marion Penninger*

Abstract

The article proposes an iterative imputation algorithm based on the EM-Algorithm and employs it to improve the education variable in the Sample of Integrated Labour Market Biographies (SIAB), an administrative panel data set provided by the Institute for Employment Research (IAB). Since the education variable in SIAB is reported for statistical reasons only, it suffers from frequent inconsistent reports and a high and increasing share of missing values. Existing imputation procedures are mainly based on heuristic rules and there is no guidance of which procedure outperforms the others. Our iterative imputation algorithm reduces the role of heuristic decision rules and estimates the most likely educational or vocational status using information based on the employee's whole employment biography. The resulting imputed education variable does not contain inconsistent reports. Furthermore, the share of missing spells is reduced by 87 percent. After imputation, the education variable shows better congruence to independent survey data (ALWA). The article focuses on the results for a (large) subgroup of SIAB (West German employees born after 1960 with a single main job). However, robustness checks reveal that the final education variable is stable with respect to different samples, termination criteria and control variables. Hence, we conclude that our imputation algorithm can serve as a blueprint for further expansions.

Zusammenfassung

Der vorliegende Artikel nutzt ein iteratives Imputations-Verfahren, das auf dem EM-Algorithmus basiert, zur Korrektur der Bildungsvariable in der Stichprobe der Integrierten Arbeitsmarktbiographien (SIAB), einem administrativen Paneldatensatz des Instituts für Arbeitsmarkt- und Berufsforschung (IAB). Die Bildungsvariable enthält einen großen Anteil an Spells, für die entweder gar kein Bildungsstatus vorliegt oder die als inkonsistent gelten müssen. Bisherige Imputationsverfahren sind größtenteils heuristischer Natur und es ist unklar, welches der Verfahren den anderen vorzuziehen ist. Unser itera-

* We are grateful to Wolfgang Biersack and Roland Weigand for helpful suggestions and valuable input. Special thanks to Johannes Ludsteck for his extraordinary help with the ALWA data set.

tives Imputationsverfahren reduziert den Einsatz von heuristischen Entscheidungsregeln und schätzt den wahrscheinlichsten Bildungsstatus. Grundlage für die Schätzungen sind erklärende Variablen, die während des gesamten Auftretens eines Beschäftigten im Datensatz gesammelt werden können. Die resultierende imputierte Bildungsvariable enthält keine inkonsistenten Bildungsverläufe mehr. Zudem verringert sich die Anzahl der Spells mit fehlenden Bildungsangaben um ca. 87 Prozent. Der Artikel setzt den Fokus auf die Ergebnisse für eine (große) Teilgruppe von SIAB (westdeutsche Beschäftigte, die nach 1960 geboren sind und nicht mehrfachbeschäftigt sind). Da Robustheitschecks eine zuverlässig hohe Stabilität der korrigierten Bildungsvariable bezüglich einer Variation der Stichprobe, der Abbruchkriterien und der Kontrollvariablen zeigen, kann unser vorgeschlagener Imputationsalgorithmus mit wenigen Modifikationen auch auf andere Untergruppen von SIAB ausgedehnt werden.

JEL Classification: C55, C81, I21, J21

1. Introduction

The use of large administrative micro data sets becomes more and more common in empirical studies. For example, the Sample of Integrated Labour Market Biographies (SIAB)[1] ranges among the most important data sources for labour market research by international scholars. The data set goes back to 1975 and includes labour market spells of individual workers on a daily basis. It is based on a variety of data sources and hence gives a holistic insight into a person's working life. The data source comprises important variables such as occupation, economic sector, wage, type of transfer payments, age, sex, place of work, place of residence or education for individuals in the German labour force.

Some core variables of the SIAB data set can be considered as highly reliable. For instance, the reported earnings give rise to claims against the social insurance system and false reporting by the employer is legally sanctioned. Hence the quality of the corresponding information in the data set is very good. Some other variables in the data set are collected for statistical reasons only. There are no consequences in case of not or false reporting. Hence the information quality depends on the reliability and accuracy of the employer reporting the corresponding data. This is the case, for instance, with the employee's educational or vocational status captured in the variable *bild*. As a consequence, the quality of the information is somewhat mixed. Inconsistencies in the individual educational and vocational biographies can easily be detected. As an example, this is the case if, according to the information given by the current employer, an employee loses a once-attained educational status as reported earlier by the same or a previous employer. Since firms are to report the highest

[1] See vom Berge et al. (2013).

educational degree attained hitherto, this should not appear. Moreover, the variable suffers increasingly from a relatively high share of missing data.

The problem of inconsistencies and missing values for educational attainment information in the predecessor of SIAB, the IABS data set, has already been documented by Fitzenberger et al. (2006), Drews (2006) and Wichert and Wilke (2012). The authors find that the share of spells with missing information on education lies between 9 percent in the IABS version covering the time period from 1975 to 1997 and 11 percent in the version spanning the years from 1975 to 2001. What is more, at least 18 to 20 percent of the individuals show implausible sequences of their educational record. For the most recent version of SIAB covering the years 1975 to 2010, we find that the problem of inconsistent and missing data has even aggravated, especially during the last decade. As a matter of fact, around 30 percent of the employment spells do not contain valid information on the educational status in 2010.

Given the high relevance of the education variable for empirical labour market analysis, several suggestions have been made to improve its information content and to purge the data from evident inconsistencies. However, previous proposals suffer from several drawbacks. In general, they rely on heuristic decision rules that are disputable and leave some scope for subjective assessments. Moreover, the different procedures typically lead to different results so that the researcher has to decide which one to take. By contrast, our contribution to the literature aims at using the entire information on education and personal characteristics in the whole sequence of individual spells contained in the data set and developing an iterative imputation procedure that is data-driven and as far as possible abstains from heuristic rules. The basic idea of our imputation algorithm can be summarized as follows. We divide the information content of the educational variable in four dichotomous statuses: "vocational training completed (yes-no)", "upper secondary / high school (yes-no)", "academic education (yes-no)" and "university degree (yes-no)". For all these four categories, we first scan the individual biographies in order to identify almost sure cases either for yes or no. For example, if there are unambiguous records indicating a completed vocational training given independently by different employers and there is no contradicting record in the entire individual biography, we consider this as overwhelming evidence that the information is correct. Having identified these almost sure cases, we estimate a linear probability model using information on individual workers' characteristics that can be generated from the panel data. From the estimation results one can calculate the predicted probability that an employee has attained a certain educational degree or vocational qualification until his or her last appearance in the data set. Repeated estimation leads to improved parameter estimates from which in turn an update of the probability of educational attainment at the individual level can be obtained. The two-steps procedure is repeated until convergence. This procedure corresponds to the well-known Expectation-Maximization (EM) algorithm intro-

duced by Dempster et al. (1977). During the expectation step, the affiliation of the individual workers to the dichotomous categories is improved, whereas, during the maximization step, the parameters of the model are fitted to the most recent affiliation.

Given the results of the EM algorithm for the four dichotomous categories, we are able to unambiguously define the educational status obtained by an individual worker. With this information and the fact that a once-attained educational status cannot be lost later on, one can impute unique and consistent individual educational biographies. The results of the procedure are promising. The outcome is reasonable and robust with respect to different samples, termination criteria and control variables. The share of missing data for the educational variable can substantially be reduced, namely from 19.4 to 2.1 percent. The remaining missings are limited to cases in which there are no exploitable data in our sample with respect to the individual's educational status during the whole employment biography or from not reaching the age condition applying backward extrapolation. For more information see Section 3.

The remainder of the article is structured as follows: Section 2 introduces the SIAB data set and the main problems with respect to the education variable. In Section 3, imputation methods suggested by other authors are discussed and we present our estimation method based on an iterative imputation algorithm for improving the education variable in SIAB. Section 4 presents the results. These are compared with the outcome of alternative methods in Section 5 where we also present a comparison with survey data for a subsample. Section 6 provides some robustness checks and Section 7 concludes.

2. The Education Variable in the SIAB Data Set

2.1 Basic Description

The Sample of Integrated Labour Market Biographies (SIAB) is a large data set provided by the Institute for Employment Research (IAB) in Nuremberg, Germany[2]. It is based on a variety of data sources and hence provides detailed information about an individual's (un)employment history on the German labour market. It is used especially (though not exclusively) by labour market researchers in a variety of projects that rely on individual labour market biographies.

Basically, SIAB is a 2 percent random sample of the population collected in the *Integrated Employment Biographies* (IEB) that comprises all individuals in Germany between 1975 to 2010 who have either

[2] Sample of Integrated Labour Market Biographies (SIAB) 1975 – 2010, Nuremberg 2013. For detailed information on SIAB, see, e.g., vom Berge et al. (2013).

- been employed covered by the social insurance system,
- worked marginal part-time (since 1999),
- received unemployment benefits (since 1975) or social assistance (since 2005), or,
- been registered as job-seeking (since 2000).

This amounts to 1,639,325 individuals whose employment biographies are documented in 45,793,010 lines of data. All characteristics are – in principle – reported exactly to the day which allows profound analyses of an individual's professional advancements and declines as well as in-depth research on the consistency of any labour market biography as explained in Section 3.

According to Bender et al. (2014), SIAB has been used by at least 117 new research projects over the past three years. The usage of the data set is likely to be extended in the future because of an improved data access. The weakly anonymized version of the SIAB, which is the basis for our imputation procedure, can be accessed in principle at a variety of of domestic and foreign sites outside the Research Data Centre in Nuremberg, e.g., via remote data processing.[3]

In the SIAB data set, the variable *bild* contains the employee's highest qualification achieved hitherto reported by the employers in the employment notification procedure.[4] It is a combined variable used to collect information on both school education and vocational qualifications of each employee. The variable *bild* can take on the following values:

- No (vocational) degree (ND): Primary school, lower secondary school, intermediate school leaving certificate or equivalent school education, all without any vocational qualification;
- Vocational training degree (VT): Primary school, lower secondary school, intermediate school leaving certificate or equivalent school education, all with a vocational qualification;
- High school degree (HS): With upper secondary school leaving certificate (*Abitur*), without a vocational qualification;
- High school degree and completed vocational training (HSVT): With upper secondary school leaving certificate and a vocational qualification;

[3] In 2014, the IAB increased the number of access points especially in the USA, where products of the RDC are accessible for research purposes at the University of Michigan in Ann Arbor, at Cornell University in Ithaca, NY, at the University of California in Berkeley at Harvard University in Cambridge, Massachusetts and at Princeton University, NJ.

[4] The variable contains also information on school education or vocational training degrees – with slightly different categories – stemming for the job search process from 1997 onwards.

- Technical college degree (TC): Degree from a university of applied sciences;
- University degree (UD).

Table 1 shows the distribution of categories of the original education variable in SIAB. The relative frequencies are calculated counting all spells from 1975 to 2010 of employees with a single job and excluding persons whose first appearance in the data was in East Germany and individuals born before 1960.[5] Moreover, we use data stemming from the employment history of workers (*BeH*) only.[6] These limitations reduce the total number of spells to 10,567,505.

Table 1

Shares of Qualification Categories in the Original Education Variable

Education Variable	Shares in percent	
	Unweighted	Weighted
bild = 1: No vocational training (ND)	23.80	22.89
bild = 2: Vocational training (VT)	44.52	49.98
bild = 3: High school degree (HS)	3.15	2.59
bild = 4: HS and VT (HSVT)	3.60	4.16
bild = 5: Technical college degree (TC)	2.12	2.55
bild = 6: University degree (UD)	3.68	4.15
Missing spells	19.13	13.69

Notes: The relative frequencies are calculated counting all employment spells from 1975 to 2010 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany and individuals born before 1960. The duration of the spells was used as weighting variable in the second column.

[5] The reason for excluding persons whose first appearance in the data was in East Germany is that for this group the labour market biographies are truncated which makes it impossible to track and assess their educational status before reunification. Employment biographies for persons born before 1960 may be left-censored in the sense that we – depending on the educational category and the age of the labour market entry – may not observe important educational information at the beginning of the employment history like reports on vocational training. Nonetheless, these restrictions can be lifted if some assumptions on how to deal with the excluded subgroups are made. Then it is straightforward to extend our imputation algorithm to a larger sample.

[6] It would be possible to use information on education stemming from the job search procedure like Kruppe et al. (2014) as well. But information from this data source are only included from 1997 onwards with a structural break in 2006 with the change in the German job search procedure. Therefore, vom Berge et al. (2013) recommend to abstain from a quantitative analysis of the variable *bild* stemming from job search information after 2005. As our goal is to present an imputed educational variable over the whole time span 1975 to 2010 we did not use the corresponding information.

The second column in Table 1 shows the respective frequencies in percent ignoring the duration of the spells. The third column displays the respective duration-weighted frequencies. The highest deviations occur with respect to VT and missing spells. Obviously, the duration of VT-reports is above and that of missing reports below average. The extraordinary high share of missing spells will be discussed in the next subsection.

2.2 Main Problems of the Educational Attainment Variable

As Fitzenberger et al. (2006), for instance, point out, the quality of the data in SIAB (or, IABS, respectively) can be considered high for variables such as the date and length of the episodes, the individual's age and sex, wage payments or the type of transfer payments. Since the reports of wages give rise to claims against the social insurance system, misreporting is legally sanctioned. However, some variables are not relevant for the social insurance system but reported for statistical reasons only. Among these variables is the employee's educational or vocational status (*bild*). As a consequence, the validity of this variable depends on the reliability and accuracy of the persons who are in charge of reporting the relevant data. A potential source for misreporting is that the educational status is simply transferred from previous spells if a person is employed in the same firm for a longer period of time. Hence, advancements in the employee's educational status are often not reported until he or she starts to work in a new company (see, e.g., Meinken / Koch, 2004).

However, a change of the employer does not necessarily mean that the reported educational status is more reliable from then on. Sometimes, firms only report the educational or vocational degree required for the respective job. Since the education variable (*bild*) should reflect the employee's highest educational status attained hitherto, reports where the employee supposedly loses a once-reported educational status can be considered inconsistent. Hence, though a job change can lead to correct reports for the first time or to a confirmation of previous reports, it can also be subject to error and the source of an inconsistency in the data.

The problem of educational discontinuities and inconsistencies is also documented in the work of Fitzenberger et al. (2006). In their study on the SIAB predecessor, IABS, ranging from 1975 to 1997[7], the authors find that in more than 9 percent of the spells the education information is missing while more than 18 percent of the individuals show inconsistent sequences of reports. Drews (2006) shows that in the subsequent version ranging from 1975 to 2001, the shares of missing and inconsistent reports have increased to 11 and 20 percent, respectively.

[7] For an overview on previous versions of SIAB, see, e.g., Bender et al. (2000).

With respect to missing values, the figures are even more striking in the latest version of SIAB covering the time period from 1975 to 2010. In our sample the average share of missing data amounts to almost 20 percent (see Table 1). Furthermore, almost 28 percent of the employees show at least one inconsistency within their educational biography. A strong shortfall of the original education variable is that the missing reports are not equally distributed over time.

Figure 1 shows the share of reports in which the education information is missing over time. Whereas the unweighted share has always been below 12 percent until 1998, it has increased sharply since.[8] In 2010, 30 percent of the employment spells suffer from missing education information. Since the average duration of missing spells is below that of spells with valid reports of the educational status, the shares of missing observations is somewhat lower when weighted by spell length. But even then the share of missing information in the education variable amounts to a quarter of all employment spells in 2010. This can lead to biased research results if the missing reports are not distributed randomly

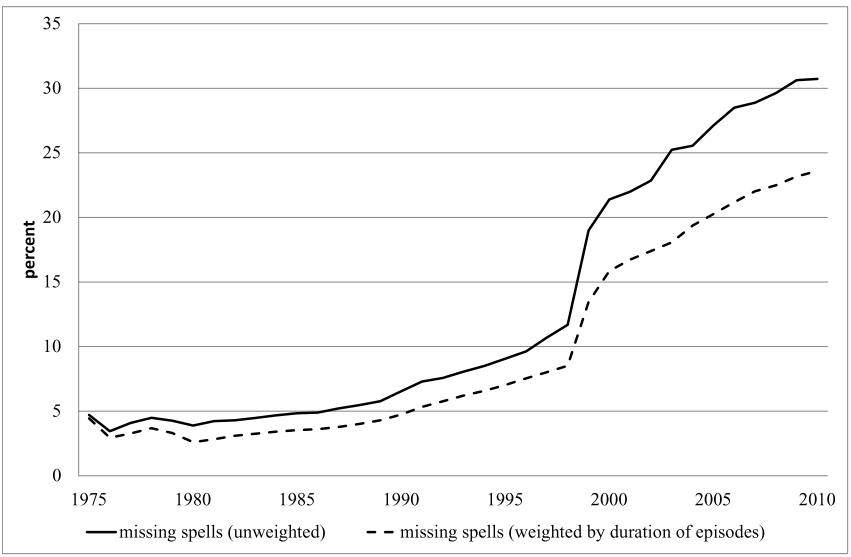


Figure 1: The share of spells with missing educational data over time

[8] To some extent, this increase might be a statistical artefact due to the inclusion of marginal part-time employment in 1999. Overlapping observations from different data sources or different forms of employment are replaced by artificial observations with new dates so that completely parallel and non-overlapping spells are created (vom Berge et al. (2013)). Hence, the probability of splitting increases with the number of different data sources. However, the duration-weighted shares should mitigate this issue.

conditional on other observed variables of interest. Furthermore, the reduced sample size due to the high share of missing information could become a relevant issue, especially in case of micro analyses for specific subgroups of workers.

3. Imputation Methods for Correcting the Educational Status

3.1 Previous Approaches to Impute the Educational Status in Administrative Data

Early approaches for imputing the variable *bild* have been developed in Bender et al. (2005) and Fitzenberger (1999).[9] However, the most comprehensive contribution and baseline for further developments comes from Fitzenberger et al. (2006) who propose three different heuristic imputation procedures to correct the education variable. The authors make use of the panel structure of the data, i.e., the possibility to exploit information on both the past and future values of the variable *bild* to infer a plausible educational status at any point in time. Imputation procedure 1 (IP1) basically extrapolates the highest educational status to subsequent spells and hence treats the problem of possible underreporting only.[10] However, if an earlier record falsely overstates the true educational status, all subsequent entries are contaminated. As a consequence, the resulting imputed education variable is expected to overreport – on average – the true educational status. The other two imputation procedures (IP2 and IP3) are more sophisticated.[11] They distinguish between invalid and valid reports and extrapolate only the latter ones to subsequent spells. While in IP2, the validity of a report depends on the frequency the same status has been reported during an employee's appearance in the data set, IP3 is based on a measure for the reliability of the firm's reporting behaviour[12] in order to assess whether a report can be considered valid or not.

The resulting imputed education variable contains substantially fewer missing data. In addition, apparent inconsistencies in the employee's educational or

---

9 For a general overview over different imputation methods, see, e.g., Rubin (2004), Little / Rubin (2002), Schafer (1997), Fellegi / Holt (1976), or Manzari (2004).

10 Fitzenberger et al. (2006) make use not only of forward extrapolation but also of backward extrapolation in case the individual enters the data set with missing education information.

11 To be precise, Fitzenberger et al. (2006) present two versions of IP2 (IP2A and IP2B). They differ insofar as IP2A has slightly stricter conditions for considering reports as reliable than IP2B.

12 This measure depends on the frequency of changes in the reported educational status of the employee by the same employer.

vocational biographies are avoided. However, there are some problems with the suggested procedure. For example, the fact that an employer changes the educational status might be a correction of a false entry and hence a signal of a certain accuracy in reporting behaviour.[13] By contrast, a false, but consistent misreporting of the educational status by the same employer due to sloppiness would be seen as a positive signal although the opposite is the case. Moreover, a practical limitation of the imputation procedure by Fitzenberger et al. (2006) is that it remains unclear which procedure out of the alternatives IP1 to IP3 outperforms the others. As a consequence, the authors recommend to conduct all different versions of their imputation procedures and to check the robustness of the results.

Drews (2006) adapts the imputation procedures of Fitzenberger et al. (2006) so that they can be employed in a later version of the data set. In addition, he uses the establishment IDs to assess the firm's reliability on reporting educational status of its employees. Using the weakly anonymized version of the data set he can take into account the reports on educational statuses of all firm's employees and not only the report of a single employee. However, the main problems connected to the heuristic nature of the imputation rules remain present.

Wichert / Wilke (2012) compare the modified educational variable using the imputation procedures of Fitzenberger et al. (2006) to educational information based on administrative data on job search. They compute misclassification probabilities and show that the measurement error decreases substantially using the edited educational variable. Furthermore, their analysis how the educational qualification or the nationality affect the probability of losing a job reveals that the correction procedure has a strong influence on the results.

Kruppe et al. (2014) test the reliability and effectivity of different correction procedures based on Fitzenberger et al. (2006), Drews (2006) and also using additional job search information. Therefore they make use of the ALWA-ADIAB data set[14] and compare the imputed educational variable in the administrative data set to the education information in the ALWA survey. They find that these heuristic correction procedures improve the quality of the educational variable. Structural biases in missreportings, however, are still present.

3.2 The Imputation Procedure

The main idea of our imputation procedure can be summarized as follows: separately for each relevant educational category in the variable *bild*, we first

[13] IP3 allows the firm to change the educational status of the employee only once to be considered reliable.

[14] For further information on ALWA-ADIAB see Section 5.

estimate the probability that the employee has achieved the respective educational or vocational degree at some point during his or her employment biography. To this end we use the entire information on education and personal characteristics in the whole sequence of individual spells contained in the data set. Estimation is done by an iterative algorithm described below. With the corresponding results we can then infer the advancements of the educational status during the individual's employment career as explained in more detail in subsection 3.3.

The following paragraphs describe the imputation algorithm in detail. We use four different models each of which aims at estimating the individual probability of having attained one of the following educational statuses: (1) HS, (2) TC or UD (STUD), (3) UD, and (4) VT. The second status (STUD) is necessary to distinguish between employees without college degree and those who attained either TC or UD. Given that an employee has attained STUD according to the second model's estimation results, we use the third model in order to identify employees with (UD) and without (TC) a university degree (see also Table 2). To be precise, we specify a linear probability model as follows:

$$\text{prob}(Y_i^j = 1|X_i^j) = c^j + X_i^j\beta^j + \varepsilon_i^j, \tag{1}$$

where $j \in (HS, STUD, UD, VT)$ and $X_i^j\beta^j = \beta_1^j X_{i1}^j + \ldots + \beta_k^j X_{ik_j}^j$. Hence, Y_i^j denotes the respective dichotomous variable and X_i^j a row vector holding individual i's values of the j-specific set of explanatory variables. Not all of the four specified models contain the same set of regressors which is why k also changes with j. While the specifications $j = HS$, $j = VT$ and $j = STUD$ are estimated using all employees in the data set, the specification $j = UD$ is limited to the group of likely college graduates as estimated for $j = STUD$.

Since both the true affiliation of the individuals to the categories of Y^j and the parameter values of the respective models are unknown, we apply a version of the EM algorithm of Dempster et al. (1977). The main advantage of the EM algorithm is its broad applicability which also includes missing value situations.[15] Furthermore, Dempster et al. (1977) prove the monotone behaviour of the underlying likelihood and the convergence of the algorithm.[16] Like the EM

[15] The idea of using the EM algorithm for dealing with inconsistent data for the education variable in SIAB has also been developed by Dlugosz (2011). The author follows a different route, however. He proposes a set of rules to identify misclassified reports and replaces them with missing values. He then derives an estimator based on the EM algorithm for incomplete data and applies it to Mincer-type wage equations. Hence, the approach does not lead to an imputed or corrected education variable.

[16] According to Wu (1983), there is an error in the proof of convergence in Dempster et al. (1977). However, the author points out that other results on the monotonicity of the

procedure, our algorithm is based on two steps: During the expectation (E) step, the affiliation of the data to the categories of Y^j is improved to better fit the model parameters. During the maximization (M) step, the parameters of the model are fitted to the most recent affiliation. The two steps are iterated until convergence.

Our algorithm requires an initial distribution of Y^j to start with. For a subgroup of individuals in the data set – for instance for those with consistent information of a certain number of different employers – we can be rather sure that a certain educational status has been attained or not. For this group of "almost sure cases" we assign starting values of either $Y^j_{i,0} = 0$ or $Y^j_{i,0} = 1$. Running Equation (1) with these starting values – using only the "almost sure cases" – yields an estimate for the parameter vector $\hat{\beta}^j_0$ that can be used to calculate the fitted values of the dependent variable $\hat{Y}^j_{i,0}$ for the whole set of observations. Hence, the algorithm calculates (pseudo-)probabilities for a certain educational status for the initially "insecure cases" and potentially revises these for the "almost sure cases". Please note that $\hat{Y}^j_{i,0}$ can take values below 0 and above 1.

After this initial M-step the conditional probabilities of having attained the respective educational status are corrected in the subsequent E-step. The E-step can be described as follows: first, we calculate the conditional error terms $\hat{\varepsilon}^{j,1}_{i,0} := \hat{\varepsilon}^j_{i,0} | Y^j_i = 1$ and $\hat{\varepsilon}^{j,2}_{i,0} := \hat{\varepsilon}^j_{i,0} | Y^j_i = 0$, where $\hat{\varepsilon}^j_i$ is the deviation between the fitted and actual value of the dependent variable under the condition that worker i has (regime 1) or has not (regime 2) obtained the respective educational status j. Let λ_0 stand for the mean value of the initially assigned probabilities $\hat{Y}^j_{i,0}$ and $f^1(\cdot)$ and $f^2(\cdot)$ for the conditional densities of a normal distribution under the assumption that the observation is drawn from regime 1 ($Y^j_i = 1$) or regime 2 ($Y^j_i = 0$), respectively. Then the enhanced probability[17] that person i has obtained the educational status j can be calculated as:

$$Y^j_{i,1} = \frac{\lambda_0 f^1(\hat{\varepsilon}^{j,1}_{i,0})}{\lambda_0 f^1(\hat{\varepsilon}^{j,1}_{i,0}) + (1-\lambda_0) f^2(\hat{\varepsilon}^{j,2}_{i,0})}. \tag{2}$$

Note that for the calculation of the densities we use the mean and standard deviation from the empirical distribution of the estimated conditional error terms $\hat{\varepsilon}^{j,1}_{i,0} | \hat{Y}^j_{i,0} > 0.5$ and $\hat{\varepsilon}^{j,2}_{i,0} | \hat{Y}^j_{i,0} \leq 0.5$, respectively. After the E-step a new M-step is executed analogously to Equation (1). The procedure is repeated until convergence. As termination criterion we define that from iteration $K-1$ to iteration K the conditional probabilities $Y^j_{i,K-1}$ and $Y^j_{i,K}$ differ by less than 0.5 percentage points on average.

---

likelihood and the convergence rate are still valid and provides a correct convergence analysis.

[17] By construction, the enhanced probability is limited to values between 0 and 1.

3.3 Imputing the Educational Biography

Given the results of the iterative procedure and ensuring that a once-attained educational status cannot be lost we are able to impute unambiguous individual educational biographies. This is done by extrapolating forward all spells – again separately for Y_i^{HS}, Y_i^{STUD}, Y_i^{UD} and Y_i^{VT} – from the moment the respective status is reported first and confirmed by the estimation. In a sense, this step is similar to the forward extrapolation proposed by Fitzenberger et al. (2006). However, a crucial difference is that we do not extrapolate reports of *bild* to subsequent spells if they are not in line with the results gained during the imputation algorithms described above. For instance, if an employee is estimated to have attained a vocational training degree, but no degree from a technical college or a university during his or her appearance in the data set, we abstain from extrapolating reports of TC or UD to subsequent spells. Furthermore, we impute ND, VT, or HSVT if the employee is estimated to be not an apprentice (any more).[18]

As a result of the procedure one obtains valid information with respect to the four educational or vocational statuses in almost all spells throughout an individual's employment career. Only in cases where there are no useful educational or vocational data throughout the individual's whole labour market biography, missing data are still remaining. In order to generate the imputed education variable, we infer the respective status according to the rules described in Table 2.

Table 2

Conditions for Constructing the Categories of the Imputed Education Variable

Conditions	Value of $bild_{imp}$
$\hat{Y}_i^{HS} < 0.5; \hat{Y}_i^{STUD} < 0.5; \hat{Y}_i^{TVT} < 0.5$	1 (ND)
$\hat{Y}_i^{HS} < 0.5; \hat{Y}_i^{VT} \geq 0.5$	2 (VT)
$\hat{Y}_i^{HS} \geq 0.5; \hat{Y}_i^{VT} < 0.5$	3 (HS)
$\hat{Y}_i^{HS} \geq 0.5; \hat{Y}_i^{VT} \geq 0.5$	4 (HSVT)
$\hat{Y}_i^{STUD} \geq 0.5; \hat{Y}_i^{UD} < 0.5$	5 (TC)
$\hat{Y}_i^{STUD} \geq 0.5; \hat{Y}_i^{UD} \geq 0.5$	6 (UD)

In all cases where the individual's employment career begins with missing but continues with valid education information, we apply backward extrapola-

[18] This correction of the apprenticeship status is done by calculating occupation-specific durations and wages gained during periods of apprenticeship. Employees with extraordinary high durations of apprenticeship, wages or wage growth rates are estimated to be no apprentices (any more).

tion in order to generate the final version of $bild_{imp}$, i.e., we extrapolate the first spell with valid (imputed) education information backward up to certain age limits. In contrast to Fitzenberger et al. (2006), we replace missing education up to age limits usually needed for achieving a certain degree: 18 years (VT), 19 years (HS), 21 years (HSVT), 22 years (TC) and 24 years (UD).

The main advantage of employing our algorithm is that the imputation process is data driven and substantially reduces the use of heuristic decision rules. After convergence is achieved, the final distribution of $bild_{imp}$ reflects consistent estimates of the employees' educational and vocational statuses attained during their employment biography. By contrast, the heuristic imputation approaches proposed by Fitzenberger et al. (2006) are expected to either overstate or understate – on average – the true educational status of the employee. As a consequence, the authors recommend to conduct all different versions of their imputation procedures and to check whether the results change substantially. This is not necessary here since there is only one final version of the imputed education variable ($bild_{imp}$). Moreover, an important byproduct of our imputation algorithm are the estimated probabilities of having a certain educational status which could be used to get an impression of the size of imprecision and variance in further research.

4. Results of the Iterative Imputation Procedure

4.1 High School Degree (HS – *Abitur*)

Starting point of the iterative estimation procedure is the identification of employees whose educational status can be considered undisputed. For a quasi secure case $Y^{HS} = 1$ we require that at least two different companies (absolute condition) *and* more than 75 percent of the companies (relative condition) have reported HS during the individual's employment career. Alternatively, the condition would also be considered fulfilled if no educational statuses lower than HS were reported. By contrast, we assume $Y^{HS} = 0$ if there have never been reports of HS or higher throughout the individual's whole employment career. In general, i.e., for all $j \in (HS, STUD, UD, VT)$ and both for $Y^{j}_{i,0} = 1$ and $Y^{j}_{i,0} = 0$, the status "secure case" requires an individual's account to have reports of at least two different firms.

In total, there are 612,975 employees in the data set (see Table 4). Among them, there are 357,064 individuals (58 percent) whose educational status can be considered secure with respect to HS. Approximately 42 percent are insecure cases.

For all $j \in$ (HS,STUD,UD,VT), we make use of a consistent set of explanatory variables.[19] In case of $j =$ HS, we use the share and number of firms that

report a certain educational status (ShareND to ShareAppHS and NumberND to NumberAppHS), the year of the report (YEAR), the employee's age (AGE), dummies for sex (FEM) and foreign nationals (FOR) and categorical variables capturing the individual's age when he or she first entered the data set (FirstAge1 to FirstAge6). A constant is always included in the regression equations. Table 3 describes in detail the meaning of all explanatory variables.

Table 3

The Explanatory Variables

Variable	Description
	Relative measures
ShareND	Share of firms reporting ND
ShareVT	Share of firms reporting VT
ShareHS	Share of firms reporting HS
ShareHSVT	Share of firms reporting HSVT
ShareTC	Share of firms reporting TC
ShareUD	Share of firms reporting UD
ShareMIS	Share of firms not reporting educational information
ShareAppND	Share of firms reporting apprenticeship and ND
ShareAppHS	Share of firms reporting apprenticeship and HS
	Absolute measures
NumberND	Number of firms reporting ND
NumberVT	Number of firms reporting VT
NumberHS	Number of firms reporting HS
NumberHSVT	Number of firms reporting HSVT
NumberTC	Number of firms reporting TC
NumberUD	Number of firms reporting UD
NumberMIS	Number of firms not reporting educational information
NumberAppND	Number of firms reporting apprenticeship and ND
NumberAppHS	Number of firms reporting apprenticeship and HS
	Time and age variables
YEAR	Year of the person's last appearance in the data set
AGE	Age at the person's last appearance in the data set

Continued next page

[19] Since the imputed education information $bild_{imp}$ could be used to explain wages in further research (e.g. in Mincer-type equations or Oaxaca-Blinder decompositions), we abstain from including any individual wage information in the explanatory variables X_i in order to avoid severe endogeneity problems.

Table 3 continued

Variable	Description
	(0,1)-Dummy Variables
FEM	Sex of the employee (1=female, 0=male)
FOR	Nationality at last appearance (1=foreign, 0=German)
FirstAge1	Person younger than 18 years at first appearance
FirstAge2	Person between the age of 18 and 20 at first appearance
FirstAge3	Person between the age of 21 and 23 at first appearance
FirstAge4	Person between the age of 24 and 26 at first appearance
FirstAge5	Person between the age of 27 and 29 at first appearance
FirstAge6	Person older than 29 years at first appearance
APP0	No apprenticeship reports
APP1	Apprenticeship reports of up to 1 year
APP2	Apprenticeship reports of 1 to 2 years
APP3	Apprenticeship reports of 2 to 3 years
APP4	Apprenticeship reports of more than 3 years
SCHOOL	Occupation with school-based vocational education

Table 4

High School Degree: Quasi Secure and Insecure Cases

Educational status	Frequency
$Y_{i,0}^{HS} = 1$ (quasi secure cases)	43,375
$Y_{i,0}^{HS} = 0$ (quasi secure cases)	313,689
Insecure status of $Y_{i,0}^{HS}$	255,911
Total number of persons	612,975

The second column of Table 8 shows the final estimation results of Equation (1) in case of $j = HS$. Convergence has been achieved after four iterations. All but one explanatory variables are significant at the 1 percent level. Note that ShareND and FirstAge1 have been chosen as reference categories and were omitted from the regression. The probability that an employee has attained HS increases with the number and share of firms that report HS or a higher degree and decreases with the number and share of firms that report the status of apprenticeship without HS. The standardized coefficients show that the share of firms that report a full university degree has the strongest impact on the dependent variable. *Ceteris paribus*, foreign persons and male employees are less likely to have a high school degree. The fact that the share of people with high

school degree has increased over the past decades is reflected in the positive sign of the time trend variable.

4.2 Technical College Degree or University Degree (STUD)

The second model treats the imputation of employees with and without a technical college or university degree ($j = STUD$). Analogously to the HS-case, both an absolute *and* and a relative condition have to be fulfilled for setting $Y_{i,0}^{STUD} = 1$. We assume $Y_{i,0}^{STUD} = 0$ if no employer reported TC or UD during the worker's employment career.

Table 5 shows that there are 397,749 individuals whose educational status can be considered secure with respect to STUD, whereas the share of insecure cases is lower than for $j = HS$ (35 percent). In the corresponding regression for $j = STUD$, the set of explanatory variables is identical to that used for $j = HS$.

Table 5

Technical College or University Degree: Quasi Secure and Insecure Cases

Educational status	Frequency
$Y_{i,0}^{STUD} = 1$ (quasi secure cases)	15,052
$Y_{i,0}^{STUD} = 0$ (quasi secure cases)	382,697
Insecure status of $Y_{i,0}^{STUD}$	215,226
Total number of persons	612,975

The third column of Table 8 shows the final estimation results of Equation (1) in case of $j = STUD$. Convergence has been achieved after four iterations. All but two explanatory variables are significant at the 1 percent level. The probability that an employee has attained STUD increases with the number and share of firms that report TC or UD which are also the variables with the highest impact according to the standardized coefficients. *Ceteris paribus*, foreign persons are less likely to have a college degree. The fact that the share of people with a college degree has increased over the past decades is reflected in the positive sign of YEAR.

4.3 University Degree (UD)

The aim of the third specification is to distinguish between those who have a technical college degree and those who have attained a full university degree for all employees who are estimated to be academics. After four iterations, the iteration procedure converges. According to the final outcome of the estimates

for STUD, there are 73,213 employees with TC or UD. Among these individuals, the conditions to assume $Y_{i,0}^{UD} = 1$ or $Y_{i,0}^{UD} = 0$ are analogous to the case of $j = STUD$.

Table 6 shows that there are much more individuals who are assigned to $Y_{i,0}^{UD} = 1$ than to $Y_{i,0}^{UD} = 0$. Four fifths of the likely academics are insecure cases, i.e., cases where the information content is not sufficient to take a clear decision.

Table 6

University Degree: Quasi Secure and Insecure Cases

Educational status	Frequency
$Y_{i,0}^{UD} = 1$ (quasi secure cases)	10,421
$Y_{i,0}^{UD} = 0$ (quasi secure cases)	2,576
Insecure status of $Y_{i,0}^{UD}$	60,216
Total number of persons	73,213

After eight iterations, the algorithm converges. The final estimation results of Equation (1) are shown in the fourth column of Table 8. All coefficients are significant at the 1 percent level. The probability that an employee has attained UD increases with the number and share of firms that report UD. In general, the number and share of firms that report TC or lower have a negative impact on the dependent variable. Among all (estimated) college graduates, foreign persons are more likely to have a full university degree. The negative sign of YEAR reflects the fact that the share of persons with a full university degree among all academics has decreased during the previous decades due to the emergence of technical colleges.

4.4 Vocational Training Degree (VT)

Employees with a vocational training degree account for the lion's share of the spells in SIAB. The allocation to the secure cases works analogously to $j = HS$ or $j = STUD$, i.e. we assume $Y_{i,0}^{VT} = 1$ if at least two different companies *and* at least 75 percent of the companies have reported VT or HSVT during the individual's employment career. By contrast, we assume $Y_{i,0}^{VT} = 0$ if there have only been reports of ND or HS throughout the employee's working life.

Table 7 shows that among all 612,975 employees there are 292,872 individuals whose educational status can be considered secure with respect to VT (48 percent). Hence, the share of insecure cases is 52 percent. The set of explanatory variables is analogous to that of the other regression models. The only difference is that some additional variables are included in order to control for

the total time of apprenticeship (APP0 to APP4) and for occupations with school-based vocational education (SCHOOL).

Table 7

Vocational Training: Quasi Secure and Insecure Cases

Educational status	Frequency
$Y_{i,0}^{VT} = 1$ (quasi secure cases)	247,705
$Y_{i,0}^{VT} = 0$ (quasi secure cases)	45,167
Insecure status of $Y_{i,0}^{VT}$	320,103
Total number of persons	612,975

After five iterations, the algorithm converges. The final estimation results are shown in the fifth column of Table 8. It becomes evident that employees are less likely to have a vocational training degree if the total time spent in apprenticeship is less than the regular time of 2 to 3 years.[20] Please note that all interactions of APP0 to APP4 with SCHOOL are negative with respect to the baseline category APP1xSCHOOL. This reflects the fact that for occupations with school-based vocational education, the usual development involves a short-time apprenticeship or internship ($APP1 = 1$). The negative sign on YEAR is the mirror image of the expansion of high school and college degrees.

Table 8

Regression Results

	Dependent variable: $\text{prob}(Y_i^j = 1)$			
	$j = HS$	$j = STUD$	$j = UD$	$j = VT$
Regressors	Standardized coefficients			
ShareVT	0.0090***	0.0074***	−0.1760***	0.7965***
ShareHS	0.3053***	−0.0463***	−0.0952***	0.0311***
ShareHSVT	0.2811***	0.0162***	−0.0605***	0.2864***
ShareTC	0.2366***	0.3163***	−0.5071***	0.2376***
ShareUD	0.4122***	0.5402***	0.0442***	0.4167***
ShareMIS	0.0039***	0.0057***	−0.0516***	0.8324***
ShareAppND	−0.0128***	−0.0101***	−0.0778***	0.5576***
ShareAppHS	0.2219***	−0.0214***	−0.0776***	0.2435***
NumberND	−0.0148***	−0.0116***	−0.1271***	−0.1744***

Continued next page

[20] APP3 is chosen as baseline category and omitted from the regression.

Table 8 continued

	Dependent variable: $\text{prob}(Y_i^j = 1)$			
	j = HS	*j = STUD*	*j = UD*	*j = V T*
Regressors	Standardized coefficients			
NumberVT	−0.0306***	−0.0224***	−0.0937***	0.0722***
NumberHS	0.1627***	0.1495***	0.1887***	−0.1659***
NumberHSVT	0.0816***	0.0027***	−0.0181***	0.0199***
NumberTC	0.0734***	0.2190***	−0.1752***	0.0164***
NumberUD	0.0320***	0.1786***	0.1074***	0.0376***
NumberMIS	0.0179***	0.0025***	−0.0281***	0.0080***
NumberAppND	−0.0135***	0.0004	−0.1218***	0.0411***
NumberAppHS	0.0982***	0.0456***	0.0503***	0.0233***
YEAR	0.0077***	0.0041***	−0.0339***	−0.0154***
AGE	−0.0042***	0.0131***	0.0471***	−0.0156***
FEM	0.0060***	−0.0003	0.0146***	0.0019***
FOR	−0.0223***	−0.0093***	0.0076***	0.0007
FirstAge2	0.0155***	0.0166***	0.0749***	0.0054***
FirstAge3	0.0113***	0.0196***	0.0848***	0.0164***
FirstAge4	0.0001	−0.0083***	0.0663***	0.0127***
FirstAge5	−0.0028***	−0.0248***	0.0430***	0.0124***
FirstAge6	−0.0072***	−0.0176***	0.0381***	0.0212***
APP0				−0.0534***
APP1				−0.1159***
APP2				−0.0080***
APP4				0.0020***
SCHOOL				0.1455***
APP0xSCHOOL				−0.0962***
APP2xSCHOOL				−0.0690***
APP3xSCHOOL				−0.1033***
APP4xSCHOOL				−0.0678***
Observations	612,975	612,975	73,213	612,975
Adjusted R^2	0.8626	0.8162	0.8506	0.8490

Notes: The results are the final outcome of estimating the respective Linear Probability Models of Equation (1) with the iteration procedure as described in the text. *, **, *** denote significance at the 10, 5, 1 percent level, respectively.

4.5 The Imputed Education Variable

Table 9 shows the distribution of the education variable after imputation. The differences in the distribution of the education variable before and after imputation are visualized in Figure 2. We manage to reduce the share of missing reports from 19.1 to 2.4 percent in case of unweighted spells and from 13.7 to 1.9 percent in case the observations are weighted by the duration of the

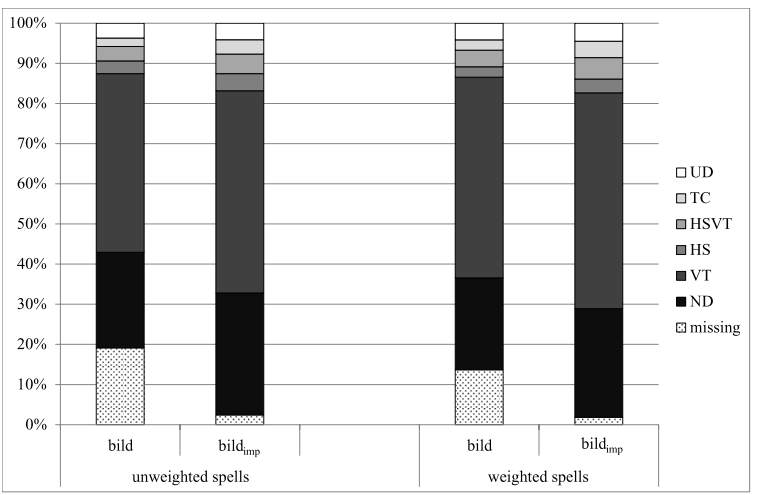
spells. The share of employees with a vocational training degree increases from 44.5 to 50.4 percent (unweighted spells). The shares of the other subgroups do not change that much in terms of percentage points. However, taking into account the low level of their shares, the changes are rather substantial, e.g., the increase from 2.1 to 3.6 percent in case of TC. This means that the number of spells with a technical college degree increases by 68 percent.

Table 9

Distribution of the Imputed Education Variable

Education ($bild_{imp}$)	Percent (unweighted)	Percent (weighted)
No vocational training (ND)	30.37	27.04
Vocational training (VT)	50.37	53.76
High school degree (HS)	4.27	3.44
HS and VT (HSVT)	4.88	5.33
Technical college degree (TC)	3.57	4.08
University degree (UD)	4.10	4.49
Missing spells	2.43	1.86

Notes: The relative frequencies are calculated using all employment spells from 1975 to 2010 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany or who were born before 1960. The duration of spells was used as weight in the weighted results.



Notes: The shares are based on all 10,567,505 observations. The duration of spells is used as weight in the weighted results.

Figure 2: Distribution of the education variable before and after imputation

Table 10 contains the imputation matrix. The values on the main diagonal show the amount of spells that remain in the same education category after imputation. Values off the main diagonal indicate changes of categories. Most of the spells with missing educational information before imputation are classified into the category ND (36 %) and VT (40 %). No assignment to an educational category is possible for 13 percent of all spells because there is no educational or vocational information throughout the individual's entire employment biography in our sample. The percentages of "stayers" based on the absolute values before imputation range from 70 percent for high school with vocational training (HSVT) to 95 percent for technical college degree (TC). In the biggest group (VT) 88 percent of the cases remain unchanged. Approximately 300.000 former ND spells change into the category VT (and vice versa). A noticeable number (148,000) of former VT spells is assigned to HSVT after the imputation procedure which leads to the conclusion that underreporting is a major issue in the original education variable.

5. Comparison with Other Imputation Methods and Data Sources

5.1 Comparison with Other Imputation Methods

One way to put the results of our imputation algorithm in context is to apply the different imputation approaches proposed by Fitzenberger et al. (2006)[21] to the recent version of SIAB spanning the years 1975 to 2010 and to compare the resulting distributions of the imputed education variables. It is important to know that IP1 to IP3 have been generated using all observations of SIAB. The figures in Table 11, however, are based on our restricted sample (10,567,505 spells) in order to guarantee comparability. The shares are not weighted by the duration of the episodes. The first column describes the distribution of the original education variable as shown in Table 1. The following three columns present the education variables after the respective imputation approaches proposed by Fitzenberger et al. (2006). The last column shows the imputed education variable resulting from our iterative algorithm.

Several points are worth noticing: all imputation procedures reduce the occurrence of missing spells substantially. IP1 results in the lowest share of missing spells and the highest shares in the educational degrees of HSVT and UD. The reason is that IP1 extrapolates reports of *bild* to all subsequent spells without assessing whether these reports are valid or made by mistake. However, as described in Subsection 3.1, IP1 is likely to *overreport* the true educational

[21] To be precise, we use the adjusted version of Drews (2006) using information on the establishment level.

Table 10

Imputation Matrix

		Imputed education variabe $bild_{imp}$							
		ND	VT	HS	HSVT	TC	UD	missing	Total
bild	ND	**2,171,741**	301,517	27,302	6,716	5,695	1,744	0	2,514,715
		86.36%	11.99%	1.09%	0.27%	0.23%	0.07%	0%	100%
	VT	297,835	**4,154,097**	25,537	148,360	59,619	18,878	0	4,704,326
		6.33%	88.30%	0.54%	3.15%	1.27%	0.40%	0%	100%
	HS	12,468	5,389	**277,481**	26,082	4,903	6,731	0	333,054
		3.74%	1.62%	83.31%	7.83%	1.47%	2.02%	0%	100%
	HSVT	2,828	49,171	23,953	**267,647**	22,016	15,088	0	380,703
		0.74%	12.92%	6.29%	70.30%	5.78%	3.96%	0%	100%
	TC	456	2,934	39	467	**213,518**	6,270	0	223,684
		0.20%	1.31%	0.02%	0.21%	95.46%	2.80%	0%	100%
	UD	1,546	4,312	356	2,228	36,501	**344,435**	0	389,378
		0.40%	1.11%	0.09%	0.57%	9.37%	88.46%	0%	100%
	Missing	722,321	805,442	96,745	64,677	34,983	40,477	**257,000**	2,021,645
		35.73%	39.84%	4.79%	3.20%	1.73%	2.00%	12.71%	100%
	Total	3,209,195	5,322,862	451,413	516,177	377,235	433,623	257,000	10,567,505
		30.37%	50.37%	4.27%	4.88%	3.57%	4.10%	2.43%	100%

status of the employee. The average educational status is somewhat lower after employing IP2 or IP3. Although Fitzenberger et al. (2006) argue that these approaches are likely to *underreport* the true educational status, our algorithm results in slightly lower shares of the statuses HSVT and UD. By contrast, the shares of HS and TC are higher compared to IP1 to IP3. The biggest difference occurs with respect to ND where our approach leads to the highest share.

Table 11

Comparison of Different Imputation Approaches

Education category	*bild*	*IP1*	*IP2*	*IP3*	$bild_{imp}$
	Shares of spells in percent				
No vocational training (ND)	23.80	24.44	26.32	23.22	30.37
Vocational training (VT)	44.52	53.83	53.33	54.16	50.37
High school degree (HS)	3.15	3.44	3.42	3.42	4.27
HS and VT (HSVT)	3.60	8.23	6.29	7.33	4.88
Technical college degree (TC)	2.12	3.00	2.59	2.77	3.57
University degree (UD)	3.68	5.11	4.58	4.92	4.10
Missing spells	19.13	1.94	3.47	4.19	2.43

Notes: bild is the original education variable. *IP*1 to *IP*3 denote the education variables after the respective imputation approaches proposed by Fitzenberger et al. (2006) which have been adapted to the recent version of SIAB spanning the years 1975 to 2010. IP2 is equivalent to IP2A in Fitzenberger et al. (2006). $bild_{imp}$ denotes the imputed education variable resulting from our iterative algorithm. The shares are based on unweighted spells.

The high similarity of the distributions of the imputed education variables could be misleading. The aggregate picture may overlie underlying dynamics on the spell level. Therefore, Table 12 compares the results of our imputation procedure ($bild_{imp}$) with the education variable using the correction procedure IP3 proposed by Fitzenberger et al. (2006). The values on the diagonal show the number of spells that are assigned to the same category by both procedures. It is worth noticing that 85 percent of all spells are assigned to identical educational statuses by both imputation procedures. However, there is some variation between the categories: The percentages range from 52 (missing) to 96 percent (ND). For some educational classes, there are noticeable differences: Almost a quarter of the spells that IP3 classifies as HSVT are in the class of VT using our imputation procedure. A similar, though not that pronounced difference also arises for VT. 12 percent of VT cases in IP3 are classified as ND using our algorithm. By contrast, almost 40 percent of the spells that are missings in IP3 are classified as ND or VT according to $bild_{imp}$. A potential reason for this outcome is that we extrapolate backward to younger ages compared to IP3 which leaves less spells with missing education information (see Section 3). To conclude,

Table 12

Comparing the Distributions for IP3 and $bild_{imp}$

		$bild_{imp}$							
		ND	VT	HS	HSVT	TC	UD	missing	Total
*IP*3	ND	**2,345,988**	83,838	7,180	2,560	2,280	577	10,925	2,453,348
		95.62%	3.42%	0.29%	0.10%	0.09%	0.02%	0.45%	100%
	VT	671,953	**4,971,053**	12,818	38,890	14,995	1,791	11,404	5,722,904
		11.74%	86.86%	0.22%	0.68%	0.26%	0.03%	0.20%	100%
	HS	26,543	2,105	**317,790**	8,609	1,357	2,416	2,516	361,336
		7.35%	0.58%	87.95%	2.38%	0.38%	0.67%	0.70%	100%
	HSVT	29,638	189,851	93,648	**449,444**	8,315	2,937	460	774,293
		3.83%	24.52%	12.09%	58.05%	1.07%	0.38%	0.06%	100%
	TC	2,328	13,746	574	2,780	**272,733**	592	211	292,964
		0.79%	4.69%	0.20%	0.95%	93.09%	0.20%	0.07%	100%
	UD	5,177	14,308	1,812	6,300	72,815	**419,627**	197	520,236
		1.00%	2.75%	0.35%	1.21%	14.00%	80.66%	0.04%	100%
	Missing	127,568	47,961	17,591	7,594	4,740	5,683	**231,287**	442,424
		28.83%	10.84%	3.98%	1.72%	1.07%	1.28%	52.28%	100%
	Total	3,209,195	5,322,862	451,413	516,177	377,235	433,623	257,000	10,567,505
		30.37%	50.37%	4.27%	4.88%	3.57%	4.10%	2.43%	100%

Table 12 emphasizes that – despite similar figures at the aggregate level – the underlying micro data show some differences between the different imputation procedures which can lead to different results in subsequent applied research.

5.2 Comparison with Information from Survey Data

Another approach to check the results of our imputation procedure is to compare it to a data set that is independent from the underlying SIAB sample. Kruppe et al. (2014) suggest using ALWA-ADIAB 7509 that combines two different parts, namely the survey information from the ALWA study (*Arbeiten und Leben im Wandel*, Working and Learning in a Changing World) and micro data from social security records. The latter stem from the same source as the SIAB data set (the IEB), contain the same variables and have the same structure as described in Section 2.1. The ALWA survey includes – among other things – comprehensive information on the educational qualification and labour market behaviour of around 10,000 individuals born between 1956 and 1988. For more information on this rich data set, see also Antoni and Seth (2011) or Antoni et al. (2011). The idea is to use the wide range of the longitudinal information on an individual's educational biography and to compare this information to the original educational variable (*bild*) and the imputed variable ($bild_{imp}$). A focus of the ALWA survey is on education and training. Hence, the answers of the surveyed persons on their educational advancements should be a valid source for a comparison to the employers' reports of the administrative data set.

As Kruppe et al. (2014) stated a comparison of the educational variables in the two parts of the ALWA-ADIAB data set is not straightforward. On the one hand, the answers in the survey part need to be recoded based on the six categories of $bild_{imp}$. On the other hand, the data generating process is completely different. The educational information in the ALWA survey is stored as different educational episodes (such as school episodes, episodes of vocational training or academic education). Hence, the ALWA data set stores qualification periods whereas the administrative ADIAB part stems from the employment notifications of the employers and contains information on *completed* educational degrees related to the respective employment episode. Therefore, episodes of completed educational achievement based on the ALWA survey are created using the ending date of the qualification episodes as starting date for newly generated episodes with the respective completed educational degree.

In the end, the imputed educational information ($bild_{imp}$) in the administrative ADIAB part[22] can be compared to that of the ALWA survey ($bild_{ALWA}$) using the same episodes.[23]

[22] In order to generate $bild_{imp}$, we use both the ADIAB data and the whole SIAB data set since our imputation algorithm requires a large sample. After the imputation procedure, the individuals not included in ALWA-ADIAB are dropped.

The results presented in Table 13 show that our procedure performs best with respect to reducing the share of missing data. Compared to $bild_{ALWA}$ – both *bild* and $bild_{imp}$ overstate the percentage of the categories ND and VT (although for the latter, $bild_{imp}$ is considerably closer to the ALWA shares). By contrast, for the categories HS and HSVT, both variables show lower frequencies compared to the education variable in ALWA. Again, $bild_{imp}$ is considerably closer to the ALWA shares than the original education variable. The shares of *IP*3 are closest to ALWA with respect to ND, HSVT and UD, whereas $bild_{imp}$ performs best in case of VT, HS and TC.

Table 13

Comparison of the Variables bild, IP3 and $bild_{imp}$ from the Administrative Data Set to Survey Information ($bild_{ALWA}$)

Education	$bild_{ALWA}$	*bild*	*IP3*	$bild_{imp}$
Missings	0.00	10.19	2.72	0.95
Among all spells with valid education information (non-missings):				
No vocational training (ND)	15.00	18.54	15.33	18.55
Vocational training (VT)	50.88	59.75	58.52	56.81
High school degree (HS)	5.87	3.30	3.10	3.97
HS and VT (HSVT)	13.64	6.56	9.86	7.52
Technical college degree (TC)	5.34	4.53	4.61	5.93
University degree (UD)	9.27	7.31	8.58	7.23

Notes: $bild_{ALWA}$ is the educational information from the ALWA survey. *bild* is the original education variable in the administrative ADIAB part. $bild_{imp}$ denotes the imputed education variable resulting from our iterative algorithm. *IP*3 denotes one of the imputed education variables proposed by Fitzenberger et al. (2006). The shares are based on duration-weighted spells. The underlying sample includes all spells from individuals in the ALWA-ADIAB data set encompassing 1975 to 2009 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany and individuals born before 1960. The analysis is based on 5415 individuals.

In order to assess *bild* and $bild_{imp}$ compared to the educational information in ALWA, it is worth looking at the share of episodes in which the educational information in *bild* and $bild_{imp}$ is different from that in $bild_{ALWA}$. Table 14 shows that in 26 percent of the spells, the education information is different from ALWA when using our imputation procedure. This is an improvement compared to the 34 percent using the original education variable. The better congruence is also reflected in the average deviation from the ALWA information. This measure is calculated as the mean absolute deviation between the

23 Kruppe et al. (2014) compare the highest educational status obtained in both data parts at a certain point in time (31/12/2006). As this may only show a small part of the picture we decided to use the whole time-span as comparison period.

educational categories of $bild_{ALWA}$ to *bild* and $bild_{imp}$, respectively, using the ordinal values one to six. For this calculation we did not take into account missing values. Using the imputed variable instead of the original educational variable reduces this distance measure from 0.48 to 0.43. Furthermore, Table 14 shows similar results for $bild_{imp}$ and *IP*3. While the share of deviating spells is marginally lower for $bild_{imp}$, the mean absolute deviation is slightly lower for *IP*3. The relatively similar performance of the imputation procedures when compared to ALWA can be explained by the fact that educational advancements usually occur earlier according to the workers' reports in ALWA[24], which is comparably impossible to detect by $bild_{imp}$ and *IP*3 if the correspondent degree has never been reported before in SIAB.

Table 14

Measuring the Deviation from $bild_{ALWA}$

Education	*bild*	*IP3*	$bild_{imp}$
Share of deviating spells	34.45%	26.67%	26.24%
Mean absolute deviation	0.4806	0.4183	0.4277

Notes: See Table 13. The mean absolute deviation is calculated without missings.

To sum up, the comparison to the educational information from ALWA shows that our imputation procedure substantially improves upon the original education variable. There still remain differences in the distributions which may stem from shortcomings of our imputation procedure or can also stem from inconsistencies in the ALWA survey.

6. Robustness Checks

In order to check the robustness of our results, we performed additional analyses. In the first alternative setting (A1 in Table 15) we include individuals born before 1960. As mentioned above we may not observe the full training or employment biographies for those born before 1960 as our data set starts with the year 1975. In order to account for the left-censoring of those employment histories, we included a dummy variable in the estimation taking the value of one for individuals born before 1960 and zero elsewhere. The coefficient of this dummy variable is negative and highly significant in all estimations except for the vocational training case where it is significantly positive.[25] This is in line

[24] These early advancements are also reflected in a higher average education level of ALWA as shown in Table 13.

[25] Due to the limitation of space the results of the four estimations for the robustness checks are not shown here but are available on request.

with the results of the imputed educational variable which are shown in the second and third column of Table 15. Note that including the individuals born before 1960 doubles the number of observations. Compared to the results of our preferred version presented in Section 4 the percentage of spells without and with vocational training is higher whereas the percentage of spells especially in the categories high school degree without and with vocational training and university degree is lower. This is also the case for the results weighted with the duration of the corresponding employment spell in the third column. The results are completely in line with the educational expansion over the last decades.

Table 15

Distributions of the Imputed Education Variable Under Alternative Underlying Samples

Education variable	Sample A1		Sample A2	
	(U)	(W)	(U)	(W)
	Shares in percent			
ND	30.50	28.00	27.16	24.50
VT	53.28	56.21	54.76	57.43
HS	2.72	2.07	3.06	2.46
HSVT	3.74	3.73	5.36	5.65
TC	3.53	3.89	3.45	3.86
UD	3.86	4.04	4.59	4.82
Missings	2.37	2.06	1.63	1.28
Number of spells	22,272,273		8,962,794	

Notes: The relative frequencies are calculated counting all employment spells from 1975 to 2010 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany. The setting of alternative 1 (A1) includes individuals born before 1960. Alternative 2 (A2) is limited to individuals born between 1960 and 1980. The results in columns 2 and 4 are unweighted (U). The duration of the spells was used as weighting variable in the third and fifth column (W).

In the second alternative setting (A2 in Table 15) the imputation procedure is limited to individuals born between 1960 and 1980. The reason is that not only the individuals born before 1960 are censored. We are also facing right-censoring of some individual employment histories. This problem might not be as severe as left-censoring since the educational history at the beginning of the employment biography such as times of vocational training is still available. However, taking only individuals born after 1960 and before 1980 might bring us as close as possible to a pure, i.e. uncensored sample. We chose the birth-cohort 1980 as boundary since these individuals are 30 years old in 2010 and have the

possibility of having completed each of the four educational categories, even a university degree, during the time span of our data set. The weighted and unweighted results are shown in column four and five of Table 15. More spells are allocated to the categories VT, HSVT, and UD, less especially to the categories ND and HS. These differences may result from the fact that in our version presented in Section 4, some individuals are allocated to the categories ND and HS but have not yet finished their educational career – which our imputation procedure cannot properly control for due to right-censoring of the data set.

So far the convergence criterion of our imputation procedure was set to 0.5 percentage points, i.e., the algorithm stops if the probability of having a certain education changes by less than 0.5 percent points on average. In order to assess to what extent our results hinge on the termination criterion, Table 16 shows the results for our imputation procedure using 1 and 0.1 percentage points as alternative termination criteria. The educational distribution is stable using 1 percentage point change on average compared to the criterion 0.5 presented in Table 9. The biggest change (around 0.4 percentage points) occurs for the category VT. Using 0.1 percentage points yields to slightly higher shares for the categories ND (around 1 percentage point) and HS and to lower shares for the other categories (especially VT). But again the distribution is relatively stable.

Table 16

Distributions of the Imputed Education Variable Under Alternative Convergence Criteria

Education variable	Convergence criterium			
	1 pp		0.1 pp	
	(U)	(W)	(U)	(W)
	Shares in percent			
ND	30.27	26.95	31.31	27.88
VT	50.75	54.09	49.68	53.21
HS	4.15	3.33	4.55	3.69
HSVT	4.91	5.36	4.72	5.20
TC	3.43	3.94	3.32	3.82
UD	4.07	4.47	3.98	4.34
Missings	2.43	1.86	2.43	1.86
Number of spells	10,567,505			

Notes: The relative frequencies are calculated counting all employment spells from 1975 to 2010 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany and individuals born before 1960. The results in columns 2 and 4 are unweighted (U). The duration of the spells was used as weighting variable in the third and fifth column (W).

For the alternative model versions presented until now we used the share of firms that report a certain educational category (for example ShareTC) during an individual's employment biography as control variables. Especially for individuals with unconventional employment biographies (e.g. vocational training before a university degree) the share of firms reporting the higher educational degree is lower than for individuals with straight biographies. This may lower the estimated probability of having a technical college or university degree. In order to account for this issue, we use alternative shares of reported technical college and university degree as control variables in the estimations for study and university degree.[26] The numerator of the alternative variable is equal to the baseline calculation. The denominator counts only the number of educational reports after the respective educational status is mentioned the first time. The results presented in Table 17 are very similar to the version in Section 4.

Table 17

Distributions of the Imputed Education Variable with Alternative Calculation of Shares

Education variable	Alternative calculation of shares (U)	(W)
	Shares in percent	
ND	30.33	27.02
VT	50.20	53.59
HS	4.27	3.44
HSVT	4.81	5.27
TC	3.54	3.98
UD	4.42	4.84
Missings	2.43	1.86
Number of spells	10,567,505	

Notes: For the estimation of the probability of having technical college and university degree we additionally included alternative shares of firms reporting technical college or university degree as control variables. For further information on this variable see text. The relative frequencies are calculated counting all employment spells from 1975 to 2010 including only employees with a single job and excluding persons whose first appearance in the data was in East Germany and individuals born before 1960. The results in column 2 are unweighted (U). The duration of the spells was used as weighting variable in the third column (W).

26 We also calculated the alternative shares for all educational categories and included them in the estimations, but the results did not change much compared to the presented version in Section 4. The results are available on request.

As expected, the shares of TC and UD are a bit higher because some individuals with college degree as highest educational category might have reports of other educational statuses beforehand.

Since the various robustness checks with different samples, termination criteria and control variables provide similar results, we conclude that our imputation procedure is rather stable.

7. Conclusions

This article aimed at improving the education variable (*bild*) in the recent version of the SIAB data set. The main problem of *bild* is the high share of unobserved data which has been increasing dramatically over the past decade. In 2010, 30 percent of the spells are missing with respect to the educational status of the employee. Even when the data are observed, a substantial share of them (almost 28 percent) shows inconsistent reports of *bild* in the course of a worker's career.

Contrary to existing approaches to impute the education variable, we propose an iterative imputation algorithm that is mainly data driven and almost completely abstains from employing heuristic rules. Furthermore, the procedure leads to a unique imputed education variable ($bild_{imp}$) which makes a calculation of several different imputation procedures expendable.

Our suggested imputation method consists of two parts. In a first step, we estimate whether an employee is likely to have attained a certain educational degree or vocational qualification during his or her appearance in the data set. Building on the results of the first step and ensuring that a once-attained educational status cannot be lost, we then impute the employee's whole educational biography in a second step.

Our imputation procedure removes approximately 87 percent of the missing values. By construction, the data set is purged from inconsistent reports. A comparison of the original and the imputed education variable reveals that the share of employees with a vocational training degree increases by about 6 percentage points. In general, the average educational status after imputation is substantially higher compared to the original variable. Taking into account the low level of the shares, the changes are rather substantial. In case of TC, for instance, the number of spells for this educational category increases by 68 percent.

Various robustness checks show that our imputation procedure is rather stable with respect to different samples, termination criteria and control variables. However, they also give an impression of the potential limitations of the proposed imputation algorithm. For instance, if there are too many workers with extreme right-censoring in the data set (i.e. they are at the beginning of

their employment career), our imputation procedure seems to encounter its limits as it is shown in Section 6.

Prospective research could benefit from expanding our sample to other subgroups of the SIAB data set. In general, it is feasible to compute the imputation procedure described in this article to workers born before 1960 (as shown in Section 6), but also to employees with more than a main job or to persons whose first appearance in the data set was in East Germany. Hence, our imputation algorithm can serve as a blueprint for further expansions. However, adjusting the sample is not straightforward and connected to potential pitfalls. For instance, workers from East Germany either suffer from severe left-censoring problems if they entered the labour market before reunification or their employment biography is rather short compared to their West German colleagues.

References

Antoni, M. / *Jacobebbinghaus*, P. / *Seth*, S. (2011): ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB 1975 – 2009 (ALWA-ADIAB 7509). FDZ-Datenreport 05 / 2011, IAB Nuremberg.

Antoni, M. / *Seth*, S. (2011): ALWA-ADIAB – Linked individual survey and administrative data for substantive and methodological research. FDZ-Methodenreport 12 / 2011, IAB Nuremberg.

Bender, S. / *Bergmann*, A. / *Fitzenberger*, B. / *Lechner*, M. / *Miquel*, R. / *Speckesser*, S. / *Wunsch*, C. (2005): Über die Wirksamkeit von Fortbildungs- und Umschulungsmaßnahmen – Ein Evaluationsversuch mit prozessproduzierten Daten aus dem IAB. Beiträge zur Arbeitsmarkt- und Berufsforschung, IAB, Nürnberg 289.

Bender, S. / *Haas*, A. / *Klose*, C. (2000): IAB Employment Subsample 1975 – 1995. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 120, 649–662.

Bender, S. / *Schmucker*, A. / *Dieterich*, I. / *Gunselmann*, I. / *Müller*, D. / *Seth*, S. / *Zakrocki*, V. (2014): FDZ-Jahresbericht 2011 – 2013. FDZ-Methodenreport: Methodische Aspekte zu Arbeitsmarktdaten.

Dempster, A. / *Laird*, N. / *Rubin*, D. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), 1 – 38.

Dlugosz, S. (2011): Give Missings a Chance – Combined Stochastic and Rule-based Approach to Improve Regression Models with Mismeasured Monotonic Covariates Without Side Information. Discussion Paper No. 11-013, Centre for European Economic Research.

Drews, N. (2006): Qualitätsverbesserung der Bildungsvariable in der IAB- Beschäftigtenstichprobe 1975 – 2001. FDZ Methodenreport. Methodische Aspekte zu Arbeitsmarktdaten. 5.

Fellegi, I. P. / *Holt*, D. (1976): A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association 71 (353), pp. 17 – 35.

Fitzenberger, B. (1999): Wages and Employment Across Skill Groups: An Analysis for West Germany. ZEW Economic Studies, Bd. 6, Heidelberg.

Fitzenberger, B. / *Osikominu*, A. / *Völter*, R. (2006): Imputation Rules to Improve the Education Variable in the IAB Employment Subsample. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 126 (3), 405 – 436.

Kruppe, T. / *Matthes*, B. / *Unger*, S. (2014): Effectiveness of data correction rules in process-produced data – The case of educational attainment. IAB-Discussion Paper 15 / 2014, Institute for Employment Research.

Little, R. J. A. / *Rubin*, D. B. (2002): Statistical Analysis with Missing Data (2nd edition). Wiley Series in Probability and Statistics, New York.

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.

Meinken, H. / *Koch*, I. (2004): BA-Beschäftigtenpanel 1998 – 2002. Codebuch. Nürnberg.

Rubin, D. B. (2004): Multiple Imputation for Nonresponse in Surveys. Wiley InterScience, New York.

Schafer, J. L. (1997): Analysis of Incomplete Multivariate Data. Chapman & Hall / CRC.

vom Berge, P. / *König*, M. / *Seth*, S. (2013): Sample of Integrated Labour Market Biographies (SIAB) 1975 – 2010. FDZ Datenreport – Documentation of labour market data.

Wichert, L. / *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.

Wu, C. F. J. (1983): On the Convergence Properties of the EM Algorithm. The Annals of Statistics 11 (1), 95 – 103.