Schmollers Jahrbuch 126 (2006), 405 – 436
Duncker & Humblot, Berlin

Imputation Rules to Improve the Education Variable in the IAB Employment Subsample*

By Bernd Fitzenberger, Aderonke Osikominu, and Robert Völter

Abstract

The education variable in the IAB employment subsample has two shortcomings: missing values and inconsistencies in the reporting rule. We propose several deductive imputation procedures to improve the variable. They mainly use the multiple education information available in the data because employees' education is reported at least once a year. We compare the improved data from the different procedures and the original data in typical applications in labor economics: educational composition of employment and wage inequality. We find that correcting the education variable shows the educational attainment of the male labor force to be higher than measured with the original data and changes some estimates of wage inequality. Our analysis does not provide a definite rule on how to choose among the different imputation procedures discussed, but we recommend correcting the original education variable.

Zusammenfassung

Die Bildungsvariable in der IAB Beschäftigtenstichprobe weist zwei Mängel auf: fehlende und der Melderegel widersprechende Werte. In dieser Arbeit werden verschiedene deduktive Imputationsverfahren entwickelt, um die Qualität der Variable zu verbessern. Die Verfahren machen sich zu Nutze, dass Bildungsinformationen für eine Person mehrfach vorhanden sind, da die Arbeitgeber mindestens einmal im Jahr eine Meldung zur Sozialversicherung abgeben müssen. Wir vergleichen die mithilfe unserer Verfahren verbesserten Daten mit den ursprünglichen Daten anhand typischer empirischer Anwendungen im Bereich der Arbeitsmarktökonomik: die Qualifikationsstruktur der Erwerbspersonen und die Lohnungleichheit. Es zeigt sich, dass die korrigierten

* This paper benefitted from the helpful comments by two anonymous referees. We gratefully acknowledge financial support by the Institut für Arbeitsmarkt- und Berufsforschung (IAB) through the research projects "Über die Wirksamkeit von FuU-Maßnahmen – Ein Evaluationsversuch mit prozessproduzierten Daten aus dem IAB (IAB project number 6-531A)" and "Die Beschäftigungswirkung der FbW-Maßnahmen 2000 – 2002 auf individueller Ebene – Eine Evaluation auf Basis der integrierten aufbereiteten IAB-Individualdatenbasis" (IAB project number 6-531.1A). We thank Stefan Bender, Michael Lechner, Ruth Miquel, Stefan Speckesser, and Conny Wunsch for very helpful discussions. All errors are our sole responsibility.

Bildungsvariablen in gestiegenen Anteilen höherer Qualifikationen resultieren. Auch das Ausmaß und die Entwicklung der Lohnungleichheit ändern sich, wenn man die korrigierten Bildungsvariablen verwendet. Unsere Studie erlaubt zwar keine Aussage darüber, welches Imputationsverfahren Idealerweise verwendet werden sollte, sie zeigt jedoch, dass es ratsam ist, die ursprüngliche Bildungsinformation in irgendeiner Form zu verbessern.

JEL Classification: C81, I21, J24, J31

Received: February 8, 2005
Accepted: March 10, 2006

1. Introduction

The IAB employment subsample (IABS) has become an important data source for empirical research on the German labor market. The IABS is a panel data set comprising administrative records for employment spells and for spells with transfer payments during periods of unemployment, see Bender et al. (2000). Compared to popular survey data sets like the German Socio-Economic Panel, the main advantages of the IABS are its large size, the long time period it covers, the almost complete absence of panel mortality, and the reliability of the core variables like date and length of spells, earnings, and type of transfer payments. However, it is well known that a number of variables are less reliable since they are not related to the purpose of the administrative reporting process producing the data. Nevertheless, research on the reliability of the IABS has been very scarce (see Fitzenberger, 1999, Steiner / Wagner, 1998, for rare exceptions). Earlier work (Cramer, 1985; Schmähl / Fachinger, 1994) pointed to problems in administrative data on employment.

The returns to education and the skill bias in labor demand are two very important issues studied in labor economics (see e.g. Card, 1999; Katz / Autor, 1999; Fitzenberger, 1999) which require a reliable measure of formal education. The IABS contains the variable *BILD* comprising information on secondary and tertiary schooling degrees as well as on completion of a vocational training degree (apprenticeship). *BILD* is based on the reports by employers and the information is extrapolated to subsequent transfer spells. This education variable exhibits a number of apparent problems. First, there is missing information for 9.52 % of the spells in the data set. Second, the education variable suffers from a large number of inconsistencies for a person over time. According to the reporting rule, employers are supposed to report the highest formal degree attained by the employee and not the degree required for the job. Hence, if a person is reported to have a certain educational degree and afterwards is reported to have a lower degree, we know that at least one of these reports must be wrong. We observe such inconsistent sequences of reports for 18.1 % of the individuals in the data set. If the incidence of these

problems is not completely at random, using the uncorrected data may result in misleading conclusions about the distribution of education and the relationship with other variables. Most of the empirical literature based on the IABS seems to use the uncorrected education variable and to exclude the observations with missings. Steiner and Wagner (1998) interpret missings in the education variable as saying that the employee exhibits no post secondary degree.

This paper develops various imputation procedures to improve the information in the IABS education variable. The main idea of our imputation approach is as follows: the panel nature of the data not only allows us to identify inconsistencies but, under reasonable assumptions, it also allows us to deduce the likely education level of a person whose education information is missing or is inconsistent for a small number of spells. If the education information is missing for a small number of spells, we impute the likely education from past or future information. If a reported degree differs for a small number of spells from the likely education, we conclude that the currently reported education is incorrect and we impute the likely education instead. Imputation has been used before to improve the education variable in the IABS. This paper extends upon the earlier work of Fitzenberger (1999, appendix) and Bender et al. (2005, chapter 3.4). We develop a number of further refinements of the basic imputation procedures. Using different versions of our imputation procedure as benchmarks, we investigate the sensitivity of empirical estimates in typical applications.

Our deductive, nonstochastic imputation rule uses the available information in the data to develop a heuristic solution to the complex problems of missing data and misclassified reports. Based on hypotheses about the reporting behavior of employers, we impute logically correct values for the actual education level, which is basically time-invariant (after an individual has reached his highest degree), when missings or inconsistencies occur. By using different hypotheses on the reporting behavior of the employers, we qualitatively evaluate the sensitivity of estimation results to the exact implementation of the imputation rule and study the statistical nature of reporting errors. A statistical validation of our imputation methods lies beyond the scope of this paper.

There exist alternative approaches in the literature which use the misclassified data directly and take misclassification into account. These methods are application-specific. Molinari (2004) makes exogenous assumptions about the misclassification probabilities and estimates identification regions for the true distribution based on the observed distribution of the misclassified data. Kane et al. (1999) estimate the returns to education when education is misclassified. The study relies on two measures of education which can both be mismeasured. These measures have to be (mean) independent of each other and of the wage conditional on true education. The latter assumption is not likely to hold

in our context. For instance, as will be discussed in detail below, if the inconsistencies in the education variable are mainly due to underreporting the level of education for people who are overqualified with respect to their position, underreporting is associated with a low wage given true education. Lewbel (2003) gives necessary assumptions to estimate average treatment effects when treatment is misclassified. Again, these conditions are unlikely to hold in our context.

Multiple imputation methods are concerned with the problem of missing data.[1] Gartner and Rässler (2005) apply multiple imputation to the problem of censored wages in the IABS. The multiple imputation framework is appropriate in a situation where the underlying (non-)response process can be modeled using the observed data. For the following reasons, an application of these methods would be very difficult in the present context. First, the variable *BILD* suffers not only from missing data but, in addition, from substantial misclassification in the non-missing data, for which the classical measurement error model is not appropriate. Second, even under the assumption that missing values are the only problem, it is questionable whether the basic condition that the occurrence of missing values is random conditional on observed quantities holds in the present context. In fact, if an employee is overqualified for his job, his employer may well underreport or not report the correct education of the employee at all because it does not match with the requirements and / or the social standing of the job. In the case of non-response, this would be correlated with the unobserved true education. Third, extensions of multiple imputation models to the panel data case can be computationally cumbersome given that we have spell data with different numbers of spells per individual.

In light of the above discussion, our deductive imputation approach has three important advantages. First, we develop a heuristic and tractable solution to a complex problem where it is very difficult to apply existing methods. Second, our method is general in the sense that we do not rely on any (conditional) independence or distributional assumptions (see discussion above on multiple imputation method proposed in the literature). Third, using different versions of our imputation procedure as benchmarks, we can investigate the sensitivity of empirical estimates in typical applications. However, this does not allow us to directly assess the statistical variability induced by the imputation.

A limitation of our approach is that it is not possible to tell which one of the different imputed education variables is the best. However, our results clearly suggest that it is advisable to use some correction of the education variable instead of ignoring the problem or resorting to *ad hoc* methods. As a practical

[1] An introduction and an overview over multiple imputation methods can be found in the textbooks of Little / Rubin (2002); Schafer (1997); or Rubin (1987).

rule, we recommend conducting the analysis based on all the imputation procedures proposed in this paper and checking whether substantive results obtained are insensitive to the imputation procedure employed.[2]

The remainder of the paper is structured as follows. Section 2 describes the IABS data and provides details on the problems concerning the education variable. Section 3 develops the different imputation procedures to improve the education variable. Section 4 examines two typical applications to compare the outcomes of the imputation procedures. Section 5 concludes. The appendix includes detailed results.

2. The IAB Employment Subsample (IABS)

2.1 Basic Description of IABS and *BILD*

We use the IABS version for the time period 1975 – 1997 distributed with detailed regional information, see Bender et al. (2000). Our imputation procedures are relevant for all versions of the IABS. The data contain daily register data for 589,825 individuals in Germany on employment spells and on spells with transfer payments from the Federal Labor Office (formerly *Bundesanstalt für Arbeit*). The IABS is a representative 1 % sample of employment. After the end of the year and when a job ends, employers have to report earnings and other socio-demographic information about their employees, such as educational degree. The earnings information and the length of the employment spells are used to calculate contributions to and benefits from the social insurance system and, hence, are very reliable. Periods of self-employment and employment as life-time civil servants (*Beamte*) that are not subject to (mandatory) social insurance are not included in the data.

The education information has to be reported with every employment spell but it bears no relevance for the social security system. To our knowledge, reporting the employee's education incorrectly has no consequences. This explains why the education variable *BILD* in the IABS is less reliable compared to information on earnings or the beginning and ending of spells. Spells on transfer payments and technical spells documenting gaps in the employment history, for instance, due to military service or maternity leave, do not provide new information on the educational level. Instead *BILD* is extrapolated during such spells based on the information in the most recent employment spell. Thus, we base our imputation procedures only on the information given in

[2] For the case of nonresponse and censoring, identification of bounds on population parameters also avoids untestable assumptions about the distribution of the missing data, see Horowitz / Manski (1998, 2000). This method is useful for analyzing 'worst-case' scenarios. It could be fruitful to explore this in future work as an alternative to our imputation approach.

employment spells. On average, the data contain 14.6 spells per person, of which 12.3 are employment spells.

Since the variable *BILD* is based on employer reports to the social security system, it is an important question how the reporting system changed between 1975 and 1997, possibly affecting the reported education. As mentioned before, the basis of the IAB employment subsample is the integrated reporting system for the social insurance, i. e., the statutory health, pension and long-term care insurance. The notification procedure was introduced in the former Federal Republic of Germany on 1 January 1973 and on 1 January 1991 – after German reunification – in the new *Länder* and East Berlin, too. Since 1973 there have been several revisions of the legislation governing the formal way in which notification has to be submitted by employers.[3] However, these changes did not concern the content – for instance the precision – of the demographic variables contained in the so-called "*Tätigkeitsschlüssel*".[4] Thus, we conclude that inconsistencies in the education variable over time are in fact attributable to employers' unreliability and not to institutional changes. This is supported by the finding that the probabilities of inconsistent and missing education reports show only very slight changes over the years (Tables 2 and 5).

The education information in the IABS distinguishes four different educational degrees (= successful completion): high school (*Abitur*), vocational training, technical college (*Fachhochschule*), and university. University is considered the highest degree, a technical college the second-highest. Since there is no clear ranking between high school and vocational training, employers have to choose among all four combinations between the two. Thus, *BILD* can take six possible meaningful values:

1. no degree at all (henceforth: ND),
2. vocational training degree (VT),
3. high school degree (HS),

---

[3] A first major revision took place in 1981 when the "*Zweite Datenerfassungsverordnung*" and the "*Zweite Datenübermittlungsverordnung*" came into effect. Their main goal was to improve the completeness of the overall amount of notifications in order to provide correct aggregate employment statistics (Wermter / Cramer, 1988). Second, in 1984 there was a change in the scope of gross earnings which are subject to social security contributions ("*Änderungsverordnung zur 2. DEVO*" (Bender et al., 1996; Fitzenberger, 1999, appendix). A third major revision of the notification procedure came into effect in 1999 ("*Datenerfassungs- und übermittlungsverordnung*"). Now, all employers are required to provide uniform information which is automatically processed.

[4] The "*Tätigkeitsschlüssel*" comprises variables that describe the job content and the qualification of the employee (cf. http://www.arbeitsagentur.de/content/de_DE/hauptstelle/a-07/importierter_inhalt/pdf/schluessel.pdf). The new "*Datenerfassungs- und übermittlungsverordnung*" actually intended to introduce a new "*Tätigkeitsschlüssel*" which has not yet been implemented.

4. high school degree and vocational training degree (HSVT),[5]
5. technical college degree (TC), and
6. university degree (UD).

We argue that these six educational outcomes can be ranked in increasing order except that no ranking exists between the second degree, VT, and the third degree, HS. We consider the comprehensive degree HSVT to be higher than both HS and VT. Furthermore, if the employee's education is not known, it can be reported as missing. According to the reporting rule, employers are supposed to report the highest degree attained by the employee, not the degree required for the current job. As a consequence, the sequence of education records should be non-decreasing over time because one can only attain higher degrees over time, not lose them. A decreasing sequence violates the reporting rule and represents evidence for inconsistent reporting behavior. All imputation procedures developed in this paper provide a corrected education variable with consistent information over time.

2.2 Spells with Missing Education

Table 1 (in the appendix) reports the distribution of the variable *BILD* in the original data. As can be seen, 9.52 % of the spells exhibit missing education information. One might suspect that missing values are mostly a problem concerning non-employment spells and short employment spells. Therefore, we also calculate the distribution of the education variable among full-time working males in 1995 excluding apprentices and weighting the spells by their length. The weighting changes the unit of measurement from spells to person years to correspond to employment. Still a weighted share of 7.35 % has missing education information. Therefore, missing values are also a sizeable problem among employees.

Next, we investigate how the incidence of missing education information among employees is related to other observed covariates in the IABS. We estimate a probit modeling the probability of a missing education report as a function of personal characteristics.[6] The estimation is based on employment

[5] In the following, we will refer to HSVT as if it were a separate degree even though it is in fact a combination of two degrees.

[6] We thank Alexandra Spitz for providing a useful classification for occupations based on the *Alphabetisches Verzeichnis der Berufsbenennungen der Bundesanstalt für Arbeit* (cf. the manual accompanying the IABS). We adjusted the classification to the occupation information given in the regional file (*BERUF* = 1–117) as follows: (i) farmers / farm managers *BERUF* = 1, 2, (ii) service workers *BERUF* = 97, 98, 110–116, (iii) operatives / craft *BERUF* = 3–57, 78–85, (iv) sales workers *BERUF* = 70–73, (v) clerical workers *BERUF* = 74–77, 89–96, (vi) administrative, professional and technical workers *BERUF* = 58–69, 86–88, 99–109.

spells only. Table 2 displays the marginal effects on the probability of a missing report. Most of the effects are significant but they are not very large compared to an observed rate of 7.8 % of missing information. Noteworthy are a 6.1 (SE 0.1) percentage points (ppoints) higher probability of a missing report for foreigners relative to Germans, a 9.2 ppoints (SE 0.3) higher probability for part-time workers with less than half the regular hours compared to full-time salaried employees and considerable differences in reporting quality across industries. Compared to the investment goods industry, the probability of a missing report is 15.7 (SE 0.4) ppoints higher in consumer services and 10.8 (SE 0.3) ppoints higher in the main construction trade.

2.3 Changes in Education across Spells

Compared to missing information, changes in the education information reported across spells are more difficult to deal with and it is crucial to analyze the sequence of reported education records across spells. If first a high degree and afterwards a low degree is reported, we know that this sequence is inconsistent with the reporting rule, but we do not know which report is incorrect. It may be that the first one overreporting or the second one underreporting or even both are incorrect. However, we can identify whether an entire sequence is consistent, i. e. nondecreasing. In the sample, 81.9 % of the persons exhibit consistent sequences of education information while 18.10 % do not.

Example 1 shows a hypothetical but representative person (all examples in this paper show hypothetical cases) with inconsistently reported education records. Only spell 3 shows education TC but all later spells show lower education with ND or VT or missing education. We do not know if the report of TC is true, but the decrease in reported education afterwards shows that some reports violate the rule of reporting the highest attained degree. Either the report of TC itself is wrong, or, in fact, the employee obtained the degree after spell 2 and before spell 3. In the latter unlikely case, all education reports after spell 3 showing a lower degree would be incorrect. In this example, there exists a second inconsistency. The report of ND at spell 15 is lower than the report of VT at spell 14.

Some insights on the reporting behavior of employers can be gained by looking at consecutive pairs of education records for the same employee. Overall, in 91.5 % of all cases, two consecutive reports are the same.[7] But there is a sharp difference depending on which employer issued the report. If both reports are by the same employer, they coincide in 97.0 % of the cases.

[7] The descriptive statistics in this section are based on employment spells only.

SPELL	*BILD*	Education	Employer	Employed
1	1	ND	1	yes
2	1	ND	2	yes
3	5	TC	3	yes
4	1	ND	1	yes
5	1	ND	0	unemployed
6	1	ND	4	yes
7	2	VT	5	yes
8	2	VT	6	yes
9	2	VT	0	unemployed
10	2	VT	0	unemployed
11	–9	missing	7	yes
12	2	VT	8	yes
13	2	VT	0	unemployed
14	2	VT	9	yes
15	1	ND	10	yes
16	2	VT	9	yes

Example 1: Person with inconsistently reported education

However, if issued by two different employers, this rate amounts to only 63.2 %. The higher stability of reports by the same employer is to be expected for the following reasons. First, attaining a higher degree often coincides with changing employers. The second explanation is rather technical and is related to the artificial splitting of some employment spells in the IABS in order to assure data privacy. This results in two consecutive spells by the same employer with the same education information. Third, this may also indicate that employers simply replicate their previous reports, causing serial correlation of reporting errors for reports on the same employee.

Furthermore, we investigate the conditional probabilities for reported education conditional on the previous report for a given person. Such a transition matrix is calculated for reports by the same employer in Table 3 and for reports from differing employers in Table 4. The high numbers (above 93 % except for HS) on the diagonal in Table 3 confirm that the same employer is likely to repeat the report given before. When the reporting employer changes, VT still has a probability of 76.9 % to be repeated in the next record. The probability for UD to be repeated is 74.7 %. This is not surprising since VT and UD are likely to be the highest degrees people attain. The other educational outcomes are, on the contrary, reported in a less stable way with probabilities of being repeated reaching at most 55.1 %.

In Table 5, we also estimate the probability that consecutive pairs of education reports on the same employee are inconsistent, i. e., the second report is lower than the first one. Since with an inconsistent pair we do not know whether the first or the second report is wrong, but only that at least one of them must be incorrect, we consider reported characteristics both in the first and the second spell. The covariates describing the employment status have the largest coefficients. Working as a trainee (apprentice obtaining a VT) in the second spell of the consecutive pair increases the probability of an inconsistent pair by 4.9 ppoints (SE 0.09), relative to working as a salaried employee. This compares to an observed total rate of 2.1 % for all pairs. Working as a skilled worker in the first spell of the pair leads to a 3.4 ppoints (SE 0.06) higher probability of an inconsistent pair, again compared to working as as salaried employee. Industry and nationality only weakly affect the probability of inconsistent reports. This is in sharp contrast to the influence that these variables have on the probability of a missing report.

3. Imputation Procedures

This section develops three imputation procedures. All imputation procedures are based on extrapolation of degrees which we will describe first.

3.1 Extrapolation and Reporting Errors

Extrapolation of educational degrees is based on three facts:

(i) the formal education level of an individual can increase when an additional degree is attained, or stay constant, but it cannot decline,

(ii) the formal education of an individual usually remains constant once the individual has entered working life, and

(iii) employers have to report the highest attained degree.

Facts (i) and (ii) state that the education of individuals is monotonically increasing but mostly constant. Fact (iii) ensures that the employers have to report the actual education of their employees and not the education necessary for the particular job which might be lower. Thus the reported education has to be equal to the actual education and hence also to be monotonically increasing.

Extrapolating plausible education reports to later spells with lower or missing education reports, we can construct an improved education variable which is monotonically increasing. For the extrapolation of education, it is helpful to distinguish three types of reporting errors: (i) underreported education, (ii) not reported education resulting in missings, and (iii) overreported education.

Spells with underreported or not reported education can be imputed with the correct education if one extrapolates the correct education from an earlier spell. In contrast, overreported education cannot be corrected by extrapolation of a correctly reported degree because the overreported degree is higher. Even worse, if one extrapolates an overreported degree to later spells with correctly reported education the quality of the education variable deteriorates. This is because extrapolation has a ratchet effect. After extrapolation, the imputed education will monotonically increase but not go down.

The obvious challenge for every imputation rule is to detect spells which are likely overreports and not to extrapolate the information. Since we do not have exogenous information about true education but only know the reported education, we cannot compare the results of an imputation rule with the true information. Hence evaluation criteria for imputation procedures requiring the true values to be known, like those in Chambers (2001), are also not applicable.[8] Instead, we propose three imputation procedures bounding the true education in distribution from above and below. The first imputation procedure (IP1) extrapolates the highest education level ever observed, including overreports. Thus, IP1 can be viewed as an upper bound for the true education. The other two procedures, IP2 and IP3, are more conservative by only extrapolating reliable reports, thus resulting in lower bounds for the true education level. IP2 uses the frequency of the report of a specific degree as an indication of its reliability and only extrapolates degrees which are reported at least three times. IP3 assesses the reporting quality of employers and only extrapolates reports from reliable employers. Taken together, these imputation procedures provide benchmarks reflecting the range of the true education information. If substantive results do not differ between IP1 and IP2 or IP3, we argue that they basically coincide with results obtained for a correct measure of education. Then, it also seems justifiable that standard errors for the estimated quantities are not adjusted for the remaining uncertainty inherent in the imputation. The next subsections will give details of the imputation procedures.[9]

3.2 Imputation Procedure 1 (IP1)

The first imputation procedure IP1 imposes no restrictions on extrapolating degrees. Every education report and hence also every overreport can be extrapolated. We argue that this procedure is likely to impute the correct education or to overstate the true education. Since we observe several education reports per person it is quite likely that true education will eventually be

[8] Other evaluation criteria exist in the literature, see e. g., Rubin (1996).

[9] More details can be found in Fitzenberger / Osikominu / Völter (2005).

reported or that an overreport will occur. In these cases, true education will be imputed or an overreport will even be imputed to many spells due to the ratchet effect. This imputation procedure may also understate the true education of some persons. An example is the case where the true education level is never reported.

The imputation procedures are implemented in four steps. Step 1 defines which reports can be extrapolated – either all, as in IP1, or only reports deemed reliable, as in IP2 or IP3. The procedures only differ in this first step. The following three steps contain the actual extrapolation and further adjustments. Next we describe the details.

Step 1: Preparation for Extrapolation

Step 1 distinguishes valid spells with extrapolatable information and invalid other spells. IP1 uses all employer reports for extrapolation. The nonemployment spells in the IABS (benefit payment spells, interruption spells) do not carry independent education information but repeat the education information of the most recent employment spell. Hence we do not use this information but treat the spells as spells with missing information. Steps 2 and 3 will extrapolate information to these invalid spells as well as to other spells with missing information. The original data include educational degrees for persons below the age of 18 years which often seem implausible. Therefore, we first impute ND for all spells in this age range.

Step 2: Forward Extrapolation

Step 2 implements the extrapolation of degrees to later spells. The procedure covers all spells of an individual person starting with the first spell and ending with the last. It extrapolates degrees to later spells if the education information reported in these later spells is lower or missing. The extrapolation stops when a spell with a higher degree or the persons last spell is reached.

The extrapolation of education information to subsequent spells has to account for the fact that the degrees HS and VT cannot be ranked. When persons have both degrees, this has to be explicitly reported. Hence, the extrapolation rule imputes HSVT if it reaches a spell with one of the two degrees and there is another previous spell for this person with the other degree.

Step 3: Backward Extrapolation

The forward extrapolation in step 2 leaves the education information missing, when spells with missing values precede a person's first spell with a valid educational report. Since the educational degree of a person is basically constant over time, we also extrapolate the first valid educational degree back-

wards to previous spells with missing information. We do not extrapolate degrees backwards beyond degree-specific age limits, because the attainment of a certain degree implies a certain number of years of schooling. The age limits are the median ages at which the degrees are reported for the first time for persons in the data. We do not impute UD backwards below the age of 29 years, TC below 27 years, HSVT below 23 years, HS below 21 years, and VT below 20 years. If the first information reported is ND, this is imputed to all prior spells. Note that the first spells of young persons can comprise missing education values, even if these persons show non-missings values in subsequent spells.

Step 4: Additional adjustments

For persons with missing education information in all spells, we impute VT if their employment status is skilled worker (*Facharbeiter*), foreman (*Polier*) or master craftsman (*Meister*). This is justified by the fact that in almost 90 % of the cases with valid education information we observe the degree VT together with such an employment status. Therefore, we impute VT for those cases. Subsequently, we also extrapolate the imputed information VT forwards and backwards analogously to steps 2 and 3.

If persons only have employment spells with other information on employment status, and education is missing for all spells, we leave it at that.

The data contain a number of parallel spells for persons who hold two or more jobs at the same time. If the imputed education variable so far takes different values for parallel spells, we finally impute the highest education information among the parallel spells to these parallel spells.

Example 2 illustrates the implementation of IP1. The forward extrapolation (*Step 2*) extrapolates VT from spell 3 to spells 4 and 5, where the lower education level ND is reported. At spell 7, HS is reported. With VT having been reported before, we assume both degrees, HS and VT, are held and impute HSVT. HSVT is considered higher than HS and extrapolated to spell, 8 and 9. For spell 10, UD is reported. Even though it is reported only once for this person, IP1 extrapolates UD to spells 11 and 12 because IP1 extrapolates every degree. Forward extrapolation alone would leave spells 1 to 2 with missing information. Hence (*Step 3*), we extrapolate VT backwards from spell 3 to spells 1 to 2. It can be seen that the imputed sequence is consistent (i. e. non-decreasing), which by construction is the case for all imputed data. In example 2, there is no missing information left. This is not necessarily the case, especially when there is only missing information about a person.

SPELL	*BILD*	Education	IP1	IP1 Education	IP2A	IP2A Education
1	–9	missing	2	VT	2	VT
2	–9	missing	2	VT	2	VT
3	2	VT	2	VT	2	VT
4	1	ND	2	VT	2	VT
5	1	ND	2	VT	2	VT
6	2	VT	2	VT	2	VT
7	3	HS	4	HSVT	4	HSVT
8	3	HS	4	HSVT	4	HSVT
9	3	HS	4	HSVT	4	HSVT
10	6	UD	6	UD	4	HSVT
11	2	VT	6	UD	4	HSVT
12	2	VT	6	UD	4	HSVT

Example 2: IP1 and IP2A

3.3 Imputation Procedure 2 (IP2)

Imputation procedure 2 (IP2) is a conservative imputation procedure, which is likely to understate the true education by restricting extrapolation to degrees which are reported at least three times. The frequency of a report serves as a measure of its reliability. If a degree is reported repeatedly, then we assume it has a lower probability to be an overreport than if is reported only once or twice. There are occasions in which a low frequency of a report arises quite naturally without indicating a likely overreport, e. g., when two degrees are obtained within a short time period. Therefore, we implement procedure 2 in two versions, IP2A and IP2B. Procedure IP2A restricts extrapolation to degrees which are reported at least three times. Procedure IP2B is less strict. Only when an inconsistent sequence of education reports indicates reporting errors for a person extrapolation is restricted to degrees which are reported at least three times. If a person's education sequence is consistent, then IP2B extrapolates every report just as IP1.

The actual implementation of IP2A is quite similar to IP1, only step 1 differs. In step 1, IP2A distinguishes valid spells carrying information reported at least three times for this person and invalid spells carrying degrees reported less often. Only information from valid spells will be extrapolated later. Spells with invalid information are later treated as spells with missing information, meaning information will be extrapolated to them from valid spells. Analogous to IP1, nonemployment spells are treated as invalid spells.

In example 2, the difference between IP2A and IP2B lies in the treatment of the single report of UD in spell 10. Procedure IP2A takes this spell as an invalid spell because UD is reported only once. Hence, it treats spell 10 as if it contains missing information and extrapolates HSVT from spell 9. IP2B and IP1 coincide in this example. In contrast to IP2A, IP2B and IP1 take spell 10 as a valid spell and impute UD afterwards.

3.4 Imputation Procedure 3 (IP3)

Analogous to IP2, IP3 is designed as a conservative imputation procedure, likely understating true education. We do not take the frequency of a report as a sign of its reliability but try to judge the reporting quality of the reporting employer.

We consider the reporting quality of an employer as being good when he always reports the same education for an employee or changes his report only once.[10] We only extrapolate reports from good reporters. IP3 treats reports from bad reporters as missing and extrapolates reports from good reporters to these spells. The hypothesis underlying this procedure is the following. Employers do not reevaluate the educational degree of their employees every time they have to give a report but tend to copy from previous reports. Thus, the frequency of the reports as such is relatively uninformative. It is more informative if an employer changes his report about an employee. And since it is very unlikely that persons attain two (or more) new degrees while being employed with one employer, we think two (or more) changes in the report indicate bad reporting quality. As noted above, education is constant for most workers after entering the labor market.

Employers might change their reports in order to correct previous reporting errors. IP3 tries to explicitly take this into account. We allow for two types of self correction. The first type consists of errors corrected immediately: an employer changes the reported degree for only one spell and switches back immediately afterwards. In this case, we ignore the switch back and forth and the employer is still classified as reliable. The second type of self-correction concerns reliable employers. If they inconsistently change their report from a higher degree to a lower degree, we assume they always wanted to report the lower degree. If a reliable employer permanently changes to a higher degree, we interpret this as the actual attainment of the higher degree. Reports from unreliable employers are set to missing. The extrapolation and additional adjustments proceed in *Steps 2–4* as in procedures IP1 and IP2. More details can be found in Fitzenberger / Osikominu / Völter (2005).

[10] Note that our data do not allow us to identify whether different employees are employed by the same employer. We can only identify which of a person's employment spells are with the same employer. Hence the reporting quality is in fact match-specific.

4. Empirical Analysis

This section compares the corrected education data resulting from the different imputation procedures to the original data. We study (i) the education mix in employment, (ii) wage inequality between and within education skill groups, (iii) how misreports are related to earnings, and (iv) the incidence of underreports.

Our basic imputation approach is based on plausible assumptions about the reporting behavior of employers and the previous section shows the importance of missing values and inconsistencies in the education variable. Therefore, we believe that empirical results using the imputed education variable are more reliable than using the uncorrected data. If substantive empirical results do not differ considerably (measured by the economic importance of the changes) when applying the different imputation procedures, we argue that the results coincide with results obtained for a correct measure of education. If results do differ, then we cannot provide a point estimate for the quantity of interest and we suspect that the different estimates provide bounds for the point estimate based on the correct education data. We cannot account further for the statistical uncertainty inherent in our imputation.

4.1 Education Mix in Employment

Table 6 shows the education shares for the original data and for the respective imputed data resulting from procedures IP1, IP2A, IP2B, and IP3 where the shares have been calculated based on the raw spells, i. e., all unweighted spells. To assess the relevance of the imputation procedure for practical applications, Table 7 reports education shares for men working in 1995 in West Germany weighted by the spell length. The Tables show that all procedures could eliminate most of the missing values. Their share decreases from 9.5 % to 1.9 – 3.2 % of the raw spells. Considering the weighted sample, we see a similar picture at a lower level. The share of missing values decreases from 7.4 % to 1.2 – 2.1 %. The remaining missing values can be explained by two reasons: (i) persons with all education information missing and (ii) the age limits for backwards extrapolation of degrees. The imputation procedures not only reduce the share of missing information but also the share of ND and HS. The shares of the education groups VT, HSVT, TC, and UD increase for the raw spells as well as for the weighted data. Next we discuss the results for the weighted data in more detail. The by far largest increase in absolute terms concerns the category VT with an increase of 5.6 – 6.9 ppoints (added to 65.3 % initially). HSVT shows the largest increase in relative terms: + 1.1 – 2.4 ppoints (added to 2.7 % initially). Considering the higher education levels, UD gains more (+ 0.8 – 1.3 ppoints added to 5.0 % initially) than TC: 0.4 – 0.7

ppoints added to 4.0 % initially. The decrease in ND is 2.5 – 5.0 ppoints from 15.1 %. The size and change of HS is small.

The imputation procedures decrease the shares of ND and HS and result in higher educational attainment among employed workers. The share of the employees holding any degree is higher (lower share of ND) and the share of the higher educational levels (TC, UD) is higher.

Comparing the different imputation procedures, IP1 shows the strongest impact on the educational composition. IP1 results in the highest shares of the higher education categories (HSVT, TC, UD) which is to be expected since it potentially extrapolates any higher report. IP2A changes the educational composition least strongly. The resulting shares of the high education categories (HSVT, TC, UD) are the lowest. IP2B is comparable to procedure IP2A except for a lower share of missing information and a higher share of VT. IP3 gives shares which are roughly in the middle between procedure IP1 and procedure IP2A. This shows that our acceptance rules based on frequency are stricter than the acceptance rule based on the reliability of the employer.

Are the differences between the imputed data from the procedures small compared to the difference to the original data? This would imply that it is important to use an imputation rule, but not very important which one. Certainly the differences between the different imputed data are small concerning the missing values and VT. For the other categories, the differences are also not too large except for the small category of HSVT. Its share goes up from 2.7 % to between 3.8 % (IP2A) and 5.1 % (IP1). For this category, the differences between the procedures are not negligible.

Additional insights can be gained from the conditional imputation rates for the different education categories given the reported education in the original data. These imputation matrices are reported in Fitzenberger, Osikominu, and Völter (2005). All procedures impute spells containing missing information with ND in about 25 % of all cases and with VT in about 50 % of all cases. At least 73 % of the non-missing reports remain unchanged. Reports from the largest category VT are rarely changed, with the procedures leaving more than 95.6 % unchanged. Only UD reports are changed less often, more than 97.1 % of them are unchanged. HS reports exhibit the highest rate of being imputed with other information. They remain unchanged with a rate of only 73.0 – 77.1 % and, if changed, they are most likely to be imputed with HSVT in 9.9 – 18.9 % of the cases. 77.4 – 83.7 % of ND reports are unchanged with 15.7 – 21.3 % being imputed with VT. All procedures provide an upward correction of the education variable but differences exist between the procedures, see Fitzenberger / Osikominu / Völter (2005) for further details.

In the literature, the six educational categories are often aggregated into three groups: (U) without a vocational training degree [ND and HS], (M) with a vocational training degree [VT and HSVT], and (H) with a higher educa-

tional degree [TC and UD] (see for instance Fitzenberger, 1999). This makes imputations within these groups irrelevant but a considerable number of imputations takes place across the groups U, M, and H, like the imputation of VT to ND. Hence, the imputation procedures are relevant at the more aggregated level as well. But the aggregation reduces the differences concerning the educational distribution, since the small group HSVT with the largest differences is aggregated with VT.

4.2 Wage Inequality Between and Within Education Groups

Now, we investigate the impact of the imputation procedures on measures of wage inequality between and within skill groups. For illustrative purposes, we focus on wage inequality among men working full time in West Germany and only consider two years, 1984 and 1997. We aggregate the six education categories into three skill groups, U, M, and H, as described in the last subsection. Table 8 shows the 20th, the 50th and the 80th percentile of the daily wage (in German Marks / DEM) for men in 1984 and in 1997 by the skill groups U, M and H. For the high-skilled H, the 50th and the 80th percentiles cannot be calculated since wages are right-censored in the data at the social security threshold. The table shows that the percentiles of the daily wage estimated with the imputed data are in most cases several DEM lower than those calculated with the original data. In 1984, this only concerns the wage percentiles for the skill groups M and H, which are estimated 1 to 4 DEM lower with the imputed data than with the original data (originally 90 – 143 DEM). In 1997, this concerns all skill groups, the estimated daily wage percentiles are up to 9 DEM lower for the imputed data. Lower estimated wage percentiles resulting from the imputed data are consistent with our view that there are many more underreports than overreports, and that underreports are associated with employees holding degrees which employers do not consider necessary for the job. Therefore, employees with underreports earn less than employees holding the same, correctly reported degree.

As a measure of wage inequality between skill groups, we consider the difference in log daily wages between the skill groups M and U at the 50th and the 20th percentiles and between the skill groups H and M only at the 20th wage percentile due to censoring of the higher wages. The numbers are given in Table 9. Imputing the education variable has a noticeable influence on the estimates of wage inequality between the low-skilled U and the medium-skilled M at the 20th wage percentile. In 1984, the estimate is lower with 0.095 for all procedures instead of 0.118 with the original variable, whereas in 1997 the estimate is higher, i. e., 0.168 to 0.219 compared to 0.163. Since the differences go in the opposite direction, the estimated 1984 – 97 increase in wage inequality varies considerably from 0.045 to

0.073 – 0.124. In other words, one would underestimate the average yearly growth rate of between wage inequality by a factor of at least two, i. e., amounting to 0.3 % based on the original variable compared to 0.6 – 0.9 % using the corrected education variables. Note further the large differences between the imputed variables themselves in the 1997 estimates which is translated to the trend estimates. The estimated inequality measures between U and M at the 50th and between H and M at the 20th wage percentile and the respective trends are not changed in a systematic way, but there are noticeable differences as well.

Table 10 reports wage inequality within the skill groups U and M. It shows the differences in log wages between the 80th and the 50th wage percentiles as well as the differences between the 50th and the 20th percentiles. Overall, the largest impact of the imputation procedures on measured inequality can be found in 1997 for skill group U below the median: the 50 % – 20 % log wage difference is measured as 0.257 – 0.269 instead of 0.228 where the results of the different imputation procedures are quite close. The other within-group wage inequalities change less than half that much, at most by 0.014. Concerning the trend between 1984 and 1997, the largest change can also be observed for the 50 % – 20 % log wage difference for skill group U. Whereas the original data result in an increase of 0.036, the imputed show a larger increase of 0.065 – 0.086. The measured trend for U and M above the median is barely affected by the imputation procedures: the growth in the 80 % – 50 % log wage difference for M shows a slightly smaller value with 0.003 – 0.014 compared to 0.021 for the original data.

Summing up, imputation affects some measures of wage inequality, especially in the lower part of the wage distribution of the low to medium-skilled groups.

4.3 Relationship between Earnings and Misreports

Next, we analyze the relationship between the individual wage and the incidence as well as the type of misreport estimating a wage regression.[11] This is of importance since wage estimations in the spirit of Kane et al. (1999), which take misclassification explicitly into account, require conditional (mean) independence of wages and measurement error given true education. We can explore whether this assumption is likely to hold by assuming true education to be close to one of the corrected values. Then we construct a missing dummy, which is one if the education information in the original data is missing, an

[11] We also compared estimated wage regressions based on the imputed data and the original data. There were only very small differences in the estimated coefficients for education dummies. Detailed results can be found in Fitzenberger / Osikominu / Völter (2005).

overreport dummy, which is one if the report in the original data is higher, and an analogous underreport dummy. If measurement error is independent of the wage conditional on true education, the dummies for the measurement error types should have insignificant coefficients in the wage regressions with the improved data. The regression controls for being foreign, occupation and industry since the incidence of missing education information was shown above to be correlated with some of these variables.

The results can be found in Table 11. The coefficient for a missing report varies between –0.106 (0.005) and –0.120 (0.004). The coefficients for underreported education are of similar size, with values between –0.104 (0.004) and –0.116 (0.003). The coefficient on overreported education is significantly negative for IP2A with –0.181 (0.012) and IP2B with –0.211 (0.017) but insignificant for IP3. Since procedure IP1 assumes there are no overreports for persons over 17 years, there is no coefficient to estimate. If we are willing to assume that the true education is not too far from one of the imputed education variables, we can conclude that an underreported or not reported education is associated with a 10 % lower wage given the true education. A lower wage, when education is underreported, is in accordance with the hypothesis that some employers report the education required for the job, not the degree attained by the employee. They pay a wage corresponding to the lower reported education. The evidence on overreported education is not conclusive. Altogether it seems that measurement error and wages are not conditionally independent given true education. Therefore, potential alternatives to imputation suggested in the literature are not applicable here.

4.4 Underreports and Overreports

This section returns to the question of incorrect education reports. Comparing the imputed data and the original data for employment spells in West Germany, the share of underreports lies between 5.8 % (IP2A) and 8.8 % (IP1) and the share of overreports between 0.2 % (IP1) and 1.0 % (IP2A). Underreports are quantitatively as important as missing values, whereas overreports are much less frequent. For this reason and because overreports differ by construction according to the different imputation procedures, we focus on underreports in the following.

The incidence of underreports is analyzed by comparing the reported education to the imputed education from IP2A in a probit regression with the set of regressors also used when analyzing missing education reports (see Table 2). The marginal effects are reported in Table 12. As the largest effect, we find a 5.7 ppoints (0.1) higher probability of an underreport for a non-skilled worker compared to a salaried employee. If the report comes from an employer who gives only one or two reports about this employee, the prob-

ability of an underreport is 3.7 ppoints (0.1) higher than when the employer gives more than five reports. Possibly, employers who anticipate employing a person for only a short time spend less effort reporting correctly. The effect of working in the main construction trade is also quite large, with a 2.6 ppoints (0.2) higher probability than in the investment goods industry. Note that, for the probability of a missing report, the effect of this industry is four times as large (see Table 2) and, analyzing inconsistencies in Table 5, there are almost no industry effects. In contrast to what we found for missing reports, foreigners are less likely to have underreports. The results for the other imputation procedures are quite comparable, see Fitzenberger / Osikominu / Völter (2005).

5. Conclusion

The education variable in the IABS shows two apparent shortcomings: missing data and observed data which is inconsistent with the reporting rule for the variable. Based on the notion that the education variable should represent a person's highest degree, that the educational degree of a working person is fairly time-constant, and that people can only attain degrees over time but not lose them, we propose different procedures to improve the variable by deductive imputation. There is no exogenous information to validate our imputation procedures. Using plausible hypotheses about the reporting process, we argue that our basic imputation procedure is likely to overstate true education and our two other refinements are likely to understate true education. If empirical results based on the different procedures are close, we argue that the imputed education variable is basically correct. In order to evaluate the impact of imputing the education variable, we analyze the educational distribution of employment as well as wage inequality between and within skill groups.

Imputation removes more than two-thirds of the missing values. The corrected data are by construction consistent with the reporting rule. Concerning the education distribution of employment, the improvement of the data matters. All procedures give higher shares for vocational training (with or without a high school degree), technical college, and university degrees and lower shares for no degree or high school only. The resulting shares do not differ a lot in size between the procedures, except for the small category of vocational training plus high school. In most dimensions, misreporting educational degrees especially affects wage inequality measured at lower percentiles in the low and medium-skilled groups. We find, for instance, that for the unskilled the measured growth from 1984 to 1997 in the difference between the median wage and the 20th wage percentile is considerably higher compared to the original data.

Overall, our results indicate some evidence in favor of the hypothesis that underreporting of educational degrees is a more severe problem than overreporting. In fact, employers tend to report the degree required for the position rather than the highest formal qualification attained by the employee. Moreover, our findings imply that usual *ad hoc* methods of dealing with or even ignoring the data quality issues regarding the education variable may bias results, especially if the focus of the analysis lies on subpopulations where the incidence of these problems is not negligible. We have demonstrated that, by exploiting the available information in the data as well as information on the institutional context, it is possible to put so much structure on the problem as to recover, in a heuristic way, an education variable that is likely to be very close to the truth.

Our analysis does not provide a definite rule on how to choose among the different imputation procedures. However, we recommend using some correction of the education variable, as suggested in our paper, instead of the common practice in existing studies involving the use of an inconsistent education variable with a large number of missing observations. In addition, without correction, the education variable in the IABS tends to understate the educational level of the employees. In actual applications, where the education variable is crucial, we recommend using all imputation procedures suggested here. If the substantive results obtained are insensitive to the use of imputation procedure, then it is very likely that they are not affected by the remaining uncertainty about the education variable. Clearly, more research on improving the data quality in the IABS is strongly needed, in particular, since the same database is used in recent evaluations of labor market reforms in Germany.

References

Bender, S. / *Bergemann,* 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 289, IAB, Nürnberg.

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

Bender, S. / *Hilzendegen,* J. / *Rohwer,* G. / *Rudolph,* H. (1996): Die IAB-Beschäftigtenstichprobe 1975–1990, Beiträge zur Arbeitsmarkt- und Berufsforschung 197, IAB, Nürnberg.

Card, D. (1999): The Causal Effect of Education on Earnings, in: O. Ashenfelter / D. Card (eds.), Handbook of Labor Economics, Vol. 3, Amsterdam.

Chambers, R. (2001): Evaluation Criteria for Statistical Editing and Imputation, National Statistics Methodological Series No. 28.

Cramer, U. (1985): Probleme der Genauigkeit der Beschäftigtenstatistik, Allgemeines Statistisches Archiv 69, 56 – 68.

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

Fitzenberger, B. / *Osikominu,* A. / *Völter,* R. (2005): Imputation Rules to Improve the Education Variable in the IAB Employment Subsample, ZEW Discussion Paper 05 – 10, Mannheim.

Gartner, H. / *Rässler,* S. (2005): Analyzing the Changing Gender Wage Gap Based on Multiply Imputed Right Censored Wages, IAB Discussion Paper 05/2000, Nürnberg.

Horowitz, J. L. / *Manski,* C. F. (1998): Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations, Journal of Econometrics 84, 37 – 58.

Horowitz, J. L. / *Manski,* C. F. (2000): Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data, Journal of the American Statistical Association 95, 77–84.

Kane, T. J. / *Rouse,* C. E. / *Staiger,* D. (1999): Estimating Returns to Schooling when Schooling is Misreported, NBER Working Paper 7235.

Katz, L. / *Autor,* D. (1999): Changes in the Wage Structure and Earnings Inequality, in: O. Ashenfelter / D. Card (eds.), Handbook of Labor Economics, Vol. 3, Amsterdam.

Lewbel, A. (2003): Estimation of Average Treatment Effects with Misclassification, Working Paper, Boston College.

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

Mincer, J. (1974): Schooling, Experience, and Earnings, National Bureau of Economic Research, New York.

Molinari, F. (2004): Partial Identification of Probability Distributions with Misclassified Data, Working Paper, Cornell University.

Rubin, D. B. (1987): Multiple Imputation for Nonresponse in Surveys, New York.

Rubin, D. B. (1996): Multiple Imputation After 18+ Years, Journal of the American Statistical Association 91, 473 – 489.

Schafer, J. L. (1997): Analysis of Incomplete Multivariate Data, London.

Schmähl, W. / *Fachinger,* U. (1994): Prozeßproduzierte Daten als Grundlage für sozial- und verteilungspolitische Analysen – Erfahrungen mit Daten der Rentenversicherungsträger für Längschnittsanalysen, in: R. Hauser / N. Ott / G. Wagner (eds.), Mikroanalytische Grundlagen der Gesellschaftspolitik, Band 2, Erhebungsverfahren, Analysemethoden und Mikrosimulation, Berlin.

Steiner, W. / *Wagner,* K. (1998): Has Earnings Inequality in Germany Changed in the 1980's?, Zeitschrift für Wirtschafts- und Sozialwissenschaften 118 (1), 29 – 59.

Wermter, W. / *Cramer,* U. (1988): Wie hoch war der Beschäftigtenanstieg seit 1983?, Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 4, Institut für Arbeitsmarkt- und Berufsforschung, 468 – 482.

Appendix

Table 1

Distribution of the Education Variable *BILD* in the Original Data

Education (abbreviation)[a]	Coded as	Number of spells	Share of spells	Weighted share[b] male empl. 1995
Missing	–9	819,701	9.52	7.35
No vocational training degree, no high school degree (ND)	1	2,325,379	27.00	15.13
Only vocational training degree, no high school degree (VT)	2	4,794,512	55.66	65.28
Only high school degree, no vocational training degree (HS)	3	95,955	1.11	0.59
High school degree and vocational training degree (HSVT)	4	153,728	1.78	2.69
Technical college degree (TC)	5	175,603	2.04	3.97
University degree (UD)	6	249,180	2.89	4.98
Total		8,614,058	100.00	100.00

[a] In German vocational training degree means *abgeschlossene Berufsausbildung,* high school degree *Abitur,* technical college degree *Fachhochschulabschluss* and university degree *Hochschulabschluss.*

[b] Weighted Share describes the education reported for fulltime working males in West Germany in 1995. Apprentices are not included. Employment spells are weighted by their length.

Table 2

Probit Regression of Education Information Missing

Regressors	Marg. eff.	Robust SE	Regressors	Marg. eff.	Robust SE
≤ 19 years	–0.018	(0.001)**	spell ≤ 30 days	0.023	(0.001)**
30 – 39 years	0.014	(0.001)**	30 < spell ≤ 180 days	0.015	(0.000)**
40 – 49 years	0.019	(0.001)**	1 – 2 reports by empl	0.038	(0.001)**
50 – 59 years	0.021	(0.001)**	3 – 5 reports by empl	0.023	(0.001)**
60+ years	0.024	(0.002)**	year 75	0.009	(0.001)**
female	–0.002	(0.001)**	year 76	0.008	(0.001)**
married	–0.009	(0.001)**	year 77	0.005	(0.001)**
foreign	0.061	(0.001)**	year 78	0.005	(0.001)**
trainee	–0.039	(0.001)**	year 79	0.005	(0.001)**
non-skilled worker	0.040	(0.001)**	year 80	0.002	(0.001)**
skilled worker	–0.023	(0.001)**	year 81	0.001	(0.001)
master craftsman / foreman	–0.034	(0.002)**	year 82	0.001	(0.001)
home worker	0.129	(0.012)**	year 83	0.000	(0.001)
part time ≤ 18h	0.092	(0.003)**	year 84	0.000	(0.000)
part time > 18h	0.028	(0.001)**	year 86	–0.000	(0.000)
farmers / farm managers	0.011	(0.002)**	year 87	0.000	(0.001)
service workers	0.019	(0.001)**	year 88	0.001	(0.001)
sales workers	–0.016	(0.001)**	year 89	0.002	(0.001)**
clerical workers	–0.028	(0.001)**	year 90	0.005	(0.001)**
admin / profes / techn staff	–0.020	(0.001)**	year 91	0.007	(0.001)**
agriculture	0.014	(0.003)**	year 92	0.008	(0.001)**
basic industry	0.015	(0.002)**	year 93	0.010	(0.001)**
consumer goods industry	0.018	(0.002)**	year 94	0.012	(0.001)**
food industry	0.045	(0.002)**	year 95	0.013	(0.001)**
main construction trade	0.108	(0.003)**	year 96	0.013	(0.001)**
construction completion trade	0.054	(0.003)**	year 97	0.015	(0.001)**
trade	0.060	(0.002)**			
transport and communication	0.084	(0.003)**			
business services	0.090	(0.002)**			
consumer services	0.157	(0.004)**			
education, non profit org	0.022	(0.002)**			
public administration	0.026	(0.002)**			
Observed prob	0.078				
Predicted prob at $\bar{x}$	0,057				
N	6,369,039				
Pseudo R^2	0.123				

Notes: Dependent variable: dummy for reported education missing. Estimation based on all employment spells in West Germany. Base category: 20 – 29 years, male, not married, German, working fulltime as a salaried employee, occupation group operatives / craft, investment goods industry, more than five reports by the employer about the employee, spell longer than 180 days, 1985. Intercept included in estimation. Robust standard errors with clustering at the person level. * significant at 5 %, ** significant at 1 %.

Table 3

Conditional Probabilities of Education Reported Given Previous Report by the Same Employer

Education reported previously	Education reported later by the same employer						
	Missing	ND	VT	HS	HSVT	TC	UD
Missing	94.43	1.68	3.47	0.06	0.12	0.11	0.13
ND	0.35	93.73	5.76	0.06	0.07	0.02	0.01
VT	0.26	0.69	98.86	0.04	0.07	0.05	0.03
HS	0.36	1.31	4.38	87.25	5.74	0.42	0.55
HSVT	0.34	0.35	2.26	0.21	96.15	0.32	0.37
TC	0.17	0.10	0.98	0.06	0.20	98.18	0.31
UD	0.15	0.05	0.46	0.06	0.11	0.23	98.94
Total	6.87	25.51	59.27	0.98	1.90	2.36	3.10

Notes: The Table contains the conditional probabilities that the row education will be reported for a person given the previous report for the person was the column education and was reported by the same employer. Based on all employment spells.

Table 4

Conditional Probabilities of Education Reported Given Previous Report by a Different Employer

Education reported previously	Education reported later by different employer						
	Missing	ND	VT	HS	HSVT	TC	UD
Missing	35.65	25.27	34.96	0.91	1.14	0.84	1.23
ND	12.48	53.12	31.62	1.16	0.70	0.46	0.47
VT	8.75	10.81	76.91	0.49	1.45	0.92	0.68
HS	7.92	15.94	23.12	25.61	10.52	5.10	11.80
HSVT	7.45	4.19	35.13	2.23	37.91	5.57	7.51
TC	5.70	1.89	21.77	1.05	5.35	55.08	9.17
UD	4.66	0.82	9.46	1.01	4.11	5.28	74.66
Total	12.84	24.34	54.71	1.19	2.08	1.88	2.96

Notes: The Table contains the conditional probabilities that the row education will be reported for a person given the previous report for the person was the column education and was reported by a different employer. Based on all employment spells.

Table 5

Probit Regression of Inconsistent Reports

Regressors	Marg. eff.	Robust SE	Regressors	Marg. eff.	Robust SE
$\leq$ 19 years	–0.0071	(0.0001)**	food industry	–0.0006	(0.0003)
30 – 39 years	0.0005	(0.0001)**	main construction trade	0.0062	(0.0004)**
40 – 49 years	–0.0007	(0.0001)**	construction completion trade	0.0012	(0.0004)**
50 – 59 years	–0.0025	(0.0001)**	trade	–0.0002	(0.0002)
60+ years	–0.0049	(0.0002)**	transport and communication	–0.0043	(0.0002)**
female	–0.0002	(0.0001)	business services	0.0019	(0.0003)**
married	–0.0026	(0.0001)**	consumer services	–0.0001	(0.0003)
foreign	–0.0013	(0.0001)**	education, non profit org	0.0031	(0.0003)**
spell $\leq$ 30 days	0.0059	(0.0002)**	public administration	0.0025	(0.0004)**
30 < spell $\leq$ 180 days	0.0103	(0.0001)**	agriculture $(t-1)$	–0.0021	(0.0004)**
spell $\leq$ 30 days $(t-1)$	0.0087	(0.0002)**	basic industry $(t-1)$	–0.0010	(0.0003)**
30 < spell $\leq$ 180 days $(t-1)$	0.0085	(0.0001)**	consumer goods industry $(t-1)$	0.0004	(0.0003)
1– 2 reports by empl	0.0065	(0.0002)**	food industry $(t-1)$	0.0008	(0.0003)*
3 –5 reports by empl	–0.0004	(0.0001)*	main construction trade $(t-1)$	–0.0013	(0.0003)**
1–2 reports by empl $(t-1)$	0.0214	(0.0003)**	construction compl trade $(t-1)$	–0.0009	(0.0003)**
3 – 5 reports by empl $(t-1)$	0.0094	(0.0002)**	trade $(t-1)$	0.0018	(0.0002)**
trainee	0.0486	(0.0009)**	transport and comm $(t-1)$	0.0092	(0.0005)**
non-skilled worker	0.0331	(0.0006)**	business services $(t-1)$	0.0012	(0.0003)**
skilled worker	–0.0154	(0.0001)**	consumer services $(t-1)$	0.0023	(0.0004)**
master craftsman / foreman	–0.0089	(0.0001)**	education, non profit org $(t-1)$	–0.0036	(0.0002)**
home worker	0.0278	(0.0034)**	public administration $(t-1)$	–0.0038	(0.0002)**
part time $\leq$ 18h	0.0227	(0.0008)**	year 75	0.0014	(0.0005)**
part time > 18h	0.0096	(0.0004)**	year 76	–0.0007	(0.0003)**
trainee $(t-1)$	–0.0098	(0.0001)**	year 77	0.0015	(0.0003)**
non-skilled worker $(t-1)$	–0.0109	(0.0001)**	year 78	0.0030	(0.0003)**
skilled worker $(t-1)$	0.0338	(0.0006)**	year 79	0.0019	(0.0003)**
master craftsman $(t-1)$	0.0225	(0.0015)**	year 80	0.0026	(0.0003)**
home worker $(t-1)$	–0.0049	(0.0007)**	year 81	0.0022	(0.0003)**
part time $\leq$ 18h $(t-1)$	–0.0032	(0.0002)**	year 82	0.0020	(0.0003)**
part time > 18h $(t-1)$	–0.0041	(0.0002)**	year 83	0.0003	(0.0003)
farmers / farm managers	0.0021	(0.0006)**	year 84	0.0001	(0.0003)
service workers	0.0023	(0.0003)**	year 86	0.0002	(0.0003)
sales workers	–0.0056	(0.0002)**	year 87	–0.0003	(0.0003)
clerical workers	–0.0037	(0.0002)**	year 88	–0.0004	(0.0003)
admin / profes / techn staff	–0.0071	(0.0002)**	year 89	–0.0005	(0.0002)*
farmers / farm man $(t-1)$	0.0031	(0.0006)**	year 90	0.0003	(0.0003)
service workers $(t-1)$	–0.0009	(0.0002)**	year 91	0.0002	(0.0003)
sales workers $(t-1)$	0.0112	(0.0005)**	year 92	–0.0007	(0.0002)**
clerical workers $(t-1)$	0.0078	(0.0004)**	year 93	–0.0006	(0.0002)*

(Continued on next page)

Continued Table 5

Regressors	Marg. eff.	Robust SE	Regressors	Marg. eff.	Robust SE
admin / profes / techn ($t-1$)	0.0191	(0.0005)**	year 94	–0.0011	(0.0002)**
agriculture	0.0006	(0.0005)	year 95	–0.0012	(0.0002)**
basic industry	0.0010	(0.0003)**	year 96	–0.0025	(0.0002)**
consumer goods industry	0.0017	(0.0003)**	year 97	–0.0038	(0.0002)**
Observed prob	0.0213		N	5,474,652	
Predicted prob at $\bar{x}$	0.0106		Pseudo R^2	0.2092	

Notes: Dependent variable: dummy for reported education lower than in the previous report. $(t-1)$ indicates variables concerning the previous employment spell. Estimation based on all employment spells in West Germany. Base category: 20 – 29 years, male, not married, German, working fulltime as a salaried employee, occupation group operatives / craft, investment goods industry, more than five reports by the employer about the employee, spell longer than 180 days, 1985. Intercept included in estimation. Robust standard errors with clustering at the person level. * significant at 5 %, ** significant at 1 %.

Table 6

Distribution of Education Variable after Imputation, Unweighted Spells

Education	Orig. data	IP1	IP2A	IP2B	IP3
Missing	9.52	1.90	3.10	2.09	3.24
No vocational training degree, no high school degree	27.00	23.41	25.68	25.80	24.09
Only vocational training degree, no high school degree	55.66	63.78	62.13	62.89	62.77
Only high school degree, no vocational training degree	1.11	1.07	1.03	1.06	1.03
High school degree and vocational training degree	1.78	3.63	2.47	2.54	2.99
Technical college degree	2.04	2.61	2.30	2.32	2.45
University degree	2.89	3.60	3.28	3.30	3.43
Total	100.00	100.00	100.00	100.00	100.00

Notes: Shares based on all 8,614,058 spells.

Table 7

Distribution of Education Variable after Imputation, Weighted Male Employment in 1995

Education	Orig. data	IP1	IP2A	IP2B	IP3
Missing	7.35	1.18	2.09	1.28	2.01
No vocational training degree, no high school degree	15.13	10.14	12.61	12.52	11.14
Only vocational training degree, no high school degree	65.28	72.15	70.83	71.59	71.60
Only high school degree, no vocational training degree	0.59	0.46	0.54	0.55	0.52
High school degree and vocational training degree	2.69	5.05	3.76	3.83	4.21
Technical college degree	3.97	4.71	4.37	4.40	4.49
University degree	4.98	6.32	5.81	5.84	6.03
Total	100.00	100.00	100.00	100.00	100.00

Notes: The Table describes the education mix for men in West Germany working fulltime in 1995. Apprentices are not included. Spells are weighted by their length.

Table 8

Wage Percentiles for Men by Skill Group for 1984 and 1997

Year	Skill group	Percentile	Orig. data	IP1	IP2A	IP2B	IP3
1984	U	20	80	80	80	80	80
		50	97	96	97	97	97
		80	116	115	116	116	117
	M	20	90	88	88	88	88
		50	110	109	109	109	109
		80	145	142	143	143	142
	H	20	143	139	141	141	140
1997	U	20	113	107	106	108	109
		80	173	170	169	171	173
	M	20	133	129	132	130	129
		50	166	161	164	163	162
		80	222	213	217	215	214
	H	20	206	197	207	203	198

Notes: The Table contains the percentiles of the daily wages in DEM for men working fulltime in West Germany without apprentices. The skill group U comprises ND and HS, M comprises VT and HSVT; H comprises TC and UD. The 50th and the 80th wage percentile for H cannot be reported because the wage data is right censored.

Table 9

Wage Inequality for Men Between Skill Groups for 1984 and 1997

Year	Groups	At percentile	Orig. data	IP1	IP2A	IP2B	IP3
1984	M-U	50	0.126	0.127	0.117	0.117	0.117
	M-U	20	0.118	0.095	0.095	0.095	0.095
	H-M	20	0.463	0.457	0.471	0.471	0.464
1997	M-U	50	0.156	0.140	0.173	0.152	0.139
	M-U	20	0.163	0.187	0.219	0.185	0.168
	H-M	20	0.438	0.423	0.450	0.446	0.428
Change	M-U	50	0.030	0.013	0.056	0.035	0.022
	M-U	20	0.045	0.092	0.124	0.090	0.073
	H-M	20	–0.026	–0.034	–0.022	–0.026	–0.036

Notes The Table contains differences in log daily wages between skill groups at specific wage percentiles based on the wage values from Table 8.

Table 10

Wage Inequality for Men Within Skill Groups for 1984 and 1997

Year	Skill group	Measure	Orig. data	IP1	IP2A	IP2B	IP3
1984	U	50 % – 20 %	0.193	0.182	0.193	0.193	0.193
		80 % – 50 %	0.179	0.181	0.179	0.179	0.187
	M	50 % – 20 %	0.201	0.214	0.214	0.214	0.214
		80 % – 50 %	0.276	0.264	0.271	0.271	0.264
1997	U	50 % – 20 %	0.228	0.269	0.264	0.260	0.257
		80 % – 50 %	0.197	0.194	0.203	0.200	0.205
	M	50 % – 20 %	0.222	0.222	0.217	0.226	0.228
		80 % – 50 %	0.291	0.280	0.280	0.277	0.278
Change	U	50 % – 20 %	0.036	0.086	0.071	0.067	0.065
		80 % – 50 %	0.019	0.014	0.024	0.021	0.017
	M	50 % – 20 %	0.021	0.008	0.003	0.012	0.014
		80 % – 50 %	0.014	0.015	0.009	0.005	0.014

Notes: The Table contains differences in log daily wages within skill groups between the respective percentiles based on the wage values from Table 8.

Table 11

Earnings and Misreports – Mincer-type Earnings Regression (Tobit)

	IP1		IP2A		IP2B		IP3	
ND	–0.120	(0.003)	–0.120	(0.003)	–0.119	(0.003)	–0.117	(0.003)
HS	–0.032	(0.020)	–0.039	(0.018)	–0.034	(0.018)	–0.017	(0.019)
HSVT	0.085	(0.005)	0.103	(0.006)	0.101	(0.006)	0.086	(0.005)
TC	0.244	(0.005)	0.253	(0.005)	0.252	(0.005)	0.250	(0.005)
UD	0.324	(0.005)	0.340	(0.005)	0.339	(0.005)	0.328	(0.005)
age / 10	0.171	(0.005)	0.165	(0.005)	0.169	(0.005)	0.169	(0.005)
age_sq / 100	–0.012	(0.001)	–0.012	(0.001)	–0.012	(0.001)	–0.012	(0.001)
reportmiss	–0.120	(0.004)	–0.106	(0.005)	–0.112	(0.004)	–0.120	(0.004)
underreport	–0.110	(0.003)	–0.106	(0.004)	-0.104	(0.004)	–0.116	(0.003)
overreport			–0.181	(0.012)	–0.211	(0.017)	0.008	(0.020)
foreign	–0.068	(0.003)	–0.065	(0.003)	–0.067	(0.003)	–0.068	(0.003)
farmer	–0.217	(0.009)	–0.216	(0.009)	–0.217	(0.009)	–0.215	(0.009)
service worker	–0.047	(0.006)	–0.044	(0.006)	–0.047	(0.006)	–0.046	(0.006)
sales worker	0.165	(0.005)	0.165	(0.005)	0.165	(0.005)	0.165	(0.005)
clerical worker	0.229	(0.003)	0.227	(0.003)	0.227	(0.003)	0.230	(0.003)
admin worker	0.282	(0.003)	0.279	(0.003)	0.280	(0.003)	0.282	(0.003)
agriculture	–0.014	(0.005)	–0.013	(0.005)	–0.014	(0.005)	–0.013	(0.005)
basic industry	0.012	(0.003)	0.013	(0.003)	0.013	(0.003)	0.013	(0.003)
consumer goods	–0.089	(0.003)	–0.087	(0.003)	–0.088	(0.003)	–0.088	(0.003)
food industry	–0.112	(0.005)	–0.111	(0.005)	–0.111	(0.005)	–0.112	(0.005)
main construction	–0.039	(0.003)	–0.039	(0.003)	–0.038	(0.003)	–0.039	(0.003)
constr completion	–0.125	(0.004)	–0.124	(0.004)	–0.124	(0.004)	–0.123	(0.004)
trade	–0.178	(0.003)	–0.177	(0.003)	–0.177	(0.003)	–0.178	(0.003)
transport & comm	–0.140	(0.004)	–0.138	(0.004)	–0.138	(0.004)	–0.139	(0.004)
business services	–0.118	(0.004)	–0.117	(0.004)	–0.118	(0.004)	–0.119	(0.004)
consumer services	–0.359	(0.009)	–0.350	(0.009)	–0.357	(0.009)	–0.359	(0.009)
education	–0.198	(0.004)	–0.198	(0.004)	-0.199	(0.004)	–0.198	(0.004)
public admin	–0.182	(0.004)	–0.182	(0.004)	–0.182	(0.004)	–0.183	(0.004)
intercept	4.660	(0.009)	4.664	(0.009)	4.664	(0.009)		
lnsigma	–1.217	(0.004)	–1.225	(0.004)	–1.220	(0.004)	–1.218	(0.004)
N	153,431		151,769		153,199		152,258	
censored	17,302		17,228		17,294		17,162	

Notes: Dependent variable: log daily wage, which is right censored at the social security threshold. Men in West Germany working fulltime 1995, no apprentices. The omitted education is VT, omitted occupation salaried employee and omitted industry investment goods industry. Spells weighted with their length. Robust standard errors clustered at the person level are in parentheses. *reportmiss, underreport* and *overreport* are defined in comparison to the original data. No *overreport* for IP1 since this procedure assumes there are no overreports for persons over 17 years.

Table 12

Probit Regression of Underreport Compared to IP2A

Regressors	Marg. eff.	Robust SE	Regressors	Marg. eff.	Robust SE
$\leq$ 19 years	–0.048	(0.000)**	1–2 reports by empl	0.037	(0.001)**
30–39 years	0.015	(0.001)**	3–5 reports by empl	0.020	(0.001)**
40–49 years	0.007	(0.001)**	spell $\leq$ 30 days	0.016	(0.001)**
50–59 years	–0.005	(0.001)**	30 $<$ spell $\leq$ 180 days	0.011	(0.000)**
60+ years	–0.017	(0.001)**	year 75	–0.037	(0.000)**
female	–0.007	(0.001)**	year 76	–0.034	(0.000)**
married	–0.003	(0.000)**	year 77	–0.029	(0.000)**
foreign	–0.020	(0.001)**	year 78	–0.022	(0.000)**
trainee	0.029	(0.001)**	year 79	–0.016	(0.000)**
non-skilled worker	0.057	(0.001)**	year 80	–0.012	(0.000)**
skilled worker	–0.028	(0.001)**	year 81	–0.008	(0.000)**
master craftsman / foreman	–0.024	(0.001)**	year 82	–0.006	(0.000)**
home worker	0.042	(0.008)**	year 83	–0.005	(0.000)**
part time $\leq$ 18h	0.023	(0.003)**	year 84	–0.002	(0.000)**
part time $>$ 18h	0.012	(0.001)**	year 86	0.001	(0.000)*
farmers / farm managers	0.012	(0.002)**	year 87	0.002	(0.000)**
service workers	–0.002	(0.001)*	year 88	0.003	(0.000)**
sales workers	–0.007	(0.001)**	year 89	0.004	(0.001)**
clerical workers	–0.002	(0.001)	year 90	0.005	(0.001)**
admin / profes / techn staff	–0.000	(0.001)	year 91	0.005	(0.001)**
agriculture	–0.006	(0.002)**	year 92	0.005	(0.001)**
basic industry	–0.002	(0.001)	year 93	0.005	(0.001)**
consumer goods industry	0.006	(0.001)**	year 94	0.004	(0.001)**
food industry	–0.002	(0.001)	year 95	0.003	(0.001)**
main construction trade	0.026	(0.002)**	year 96	0.001	(0.001)
construction completion trade	0.002	(0.002)	year 97	–0.001	(0.001)
trade	0.006	(0.001)**			
transport and communication	–0.003	(0.001)*			
business services	0.007	(0.001)**			
consumer services	0.003	(0.001)			
education, non profit org	0.002	(0.001)			
public administration	–0.002	(0.001)			
Observed prob	0.060				
Predicted prob at $\bar{x}$	0.047				
N	6,352,330				
Pseudo R^2	0.078				

Notes: Dependent variable: dummy for reported education lower than imputed education (IP2A). Estimation based on all employment spells in West Germany. Base category: 20–29 years, male, not married, German, working fulltime as a salaried employee, occupation group operatives/craft, investment goods industry, more than five reports by the employer about the employee, spell longer than 180 days, 1985. Intercept included in estimation. Robust standard errors with clustering at the person level. * significant at 5 %, ** significant at 1 %.