Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert

Working Paper

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

ZEW Discussion Papers, No. 05-10

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2005) : Imputation Rules to Improve the Education Variable in the IAB Employment Subsample, ZEW Discussion Papers, No. 05-10, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

This Version is available at:
https://hdl.handle.net/10419/24103

Discussion Paper No. 05-10

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

Bernd Fitzenberger, Aderonke Osikominu
and Robert Völter

ZEW
Zentrum für Europäische
Wirtschaftsforschung GmbH
Centre for European
Economic Research

Discussion Paper No. 05-10

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

Bernd Fitzenberger, Aderonke Osikominu
and Robert Völter

Non-technical Summary

The IAB employment subsample (IABS) has become an important data source for empirical research on the German labor market. It consists of employment spells subject to social insurance and unemployment spells that comprise information about the beginning and the end of an (un-)employment period, daily wage, type of transfer payments, as well as several socio–demographic variables like, for instance, sex, year of birth, occupation, employment status, and education.

The reporting of the educational degree of an employee to the social insurance agencies by the employer is part of the compulsory notification system. However, the educational degree has no consequences concerning obligations or claims out of the social security neither for the employer nor for the employee. Therefore, this variable can be regarded as less reliable than other variables like wages. However, previous research on this issue has been scarce although the problem of missing and inconsistent education information is immediately apparent in the data set.

To deal with the problem of missing values and/or inconsistent sequences of educational reports, we use deductive imputation methods that make use of the panel structure of the data set. In addition, information contained in the variables employment status and age is considered when constructing the corrected education variables.

We develop three different imputation methods based on different assumptions on nature of the reporting process. Imputation procedure 1 assumes that underreports are the only possible source of inconsistencies as some employers – for reasons that are not observable to the researcher – do not report the actual educational degree of the employee but the degree required for the position. Under this scenario every higher degree is extrapolated to subsequent spells with lower or missing education information. However, as long as one cannot completely rule out the possibility of overreports this procedure is likely to induce a considerable upward bias in the corrected education variable. Therefore, we develop two further imputation procedures that distinguish between reliable reports that are used for extrapolation and unreliable reports that have to be discarded. Imputation procedure 2 assumes that the observed frequency of a reported degree can be interpreted as a sign of its reliability. Imputation procedure 3, on the contrary, is based on the assumption that reporting errors are serially correlated. In this case, a change of the reported educational degree may reveal some information on the reliability of the employer issuing the reports. Imputation procedure

1 possibly overstates education through a ratched effect when overreporting occurs. The imputation procedures 2 and 3 allow for the possibility of overreporting. They use heuristic rules to identify valid education information in a conservative way.

We argue that, if all three corrected education variables yield similar results in different applications, we are likely to have found a good approximation to true education using any of the corrected education variables.

Therefore, we compare in three typical labor economic applications the data resulting from the three imputation procedures to the original data. We find that the educational attainment of the labor force is higher than measured with the original data. Moreover, we find that some measures of between and within education group inequality, especially in the lower part of the wage distribution, are sensitive to the education variable used, whereas the estimated return to education does not vary much with the education variable used. Overall, our results indicate that there is some evidence in favor of the hypothesis that underreporting of educational degrees is a more severe problem than overreporting. The evidence suggests that, in fact, employers tend to report the degree required for the position rather than the highest qualification attained by the employee.

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

Bernd Fitzenberger[*], Aderonke Osikominu[†], Robert Völter[‡]

February 7, 2005

Abstract

The education variable in the IAB employment subsample has two shortcomings: missing values and inconsistencies with 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 the 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, wage inequality, and wage regression. We find, that correcting the education variable: (i) shows the educational attainment of the male labor force to be higher than measured with the original data, (ii) gives different values for some measures of wage inequality, and (iii) does not change the estimates in wage regressions much.

[*]Goethe University Frankfurt, ZEW, IZA, IFS. Address: Department of Economics, Goethe-University, PO Box 11 19 32 (PF 247), 60054 Frankfurt am Main, Germany. E-mail: fitzenberger@wiwi.uni-frankfurt.de

[†]Goethe University Frankfurt. E-mail: osikominu@wiwi.uni-frankfurt.de

[‡]Goethe University Frankfurt. E-mail: voelter@wiwi.uni-frankfurt.de.

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)" und "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.

1 Introduction

The IAB employment subsample (IABS) has become an important data source for empirical research on the German labor market, see Bender et al. (2000) for a recent description of the data set. The IABS is a panel data set comprising administrative records for employment spells and for spells with transfer payments during periods of unemployment. Compared to popular survey data sets like the German Socioeconomic 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, or type of transfer payments. However, it is well known among users of the IABS 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 and Wagner, 1998, for rare exceptions).[1]

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 and 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). This variable 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 degree attained by the employee and not the degree required for the job. Hence, if once in the data a person is recorded to have a certain educational degree and later is reported to have only a lower degree, we know that at least one of these reports must be wrong. If the incidence of these problems is not random, using the uncorrected data may result in misleading conclusions about the distribution of education or the relationship with other variables of interest.

[1] Earlier work (see e. g. Cramer, 1985, or Schmähl and Fachinger, 1994) pointed to some problems in the administrative data on employment subject to social security taxation.

This paper develops some 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 does not only allow 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 therein and we investigate the effects on three typical applications in labor economics.

Without knowing the unobserved true education, we cannot evaluate the validity of our imputation procedures. Hence, evaluation criteria for imputation which require the true values to be known, like those in Chambers (2001), are not applicable here. This paper proposes three imputation procedures. Based on plausible assumptions about the reporting behavior of the employers, one imputation procedure is more likely to overstate the true level of education and two are more likely to understate it. We argue that, if the resulting data from the different procedures are quite similar, our approach is likely to be close to the truth. In order to evaluate the differences across imputation procedures, we investigate three typical applications in labor economics. In our first application we analyze the education mix in employment. The second one examines changes in wage differentials across and within education groups over time. And the third one involves estimating Mincer-type earnings equations.

An advantage of imputation methods is that the resulting data can be treated as being measurement error free provided one is thinking that the imputation method applied is valid. There exist alternative approaches in the literature which use the misclassified data directly and take misclassification into account. These methods are application specific, as the following references show. Molinari (2004), in a direct misclassification approach, 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, Rouse and Staiger (1999) estimate the returns to education when education is misclassified. They rely 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 the 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.

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 three typical applications to compare the outcomes of the imputation procedures. Section 5 concludes. The appendix includes detailed results.

2 The IAB Employment Subsample (IABS)

This section first describes the IABS and the education variable *BILD*. Then, we discuss the shortcomings of *BILD*, namely, missing values and inconsistencies over time.

2.1 Basic Description of IABS and *BILD*

We use the version of the IABS (IAB-Beschäftigtenstichprobe) for the time period 1975-1997 distributed with detailed regional information (Regionalfile). A basic description of the data set can be found in Bender et al. (2000). Our imputation procedures are relevant for all versions of the IABS. The data used here contain daily register data of 589,825 individuals in Germany on their employment spells and the spells during which they receive transfer payments from the Federal Labor Office (formerly *Bundesanstalt für*

Arbeit). It is a representative 1% sample of employment subject to social security taxation and, therefore, it is not representative with respect to periods of nonemployment. After the end of the year and when a job ends, employers have to report earnings and other socio-demographic information about their employees like their educational degree to the social insurance agencies. 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 variable *BILD* in the IABS is a byproduct of a reported employment spell and bears no relevance for the social security system. To our knowledge, reporting the employee's education incorrectly has no consequences. This can explain why *BILD* 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 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 the German reunification – in the new Länder and Berlin-East, too. Since 1973 there have been several revisions of the legislation governing the formal way of how the notifications have to be submitted by the employers.[2] However, these changes did not concern the content – for in-

[2] A first major revision of the notification system took place in 1981 when the "Zweite Datenerfassungsverordnung" and the "Zweite Datenübermittlungsverordnung" came into effect. The main goal of this new formulation has been to improve the completeness of the

stance the precision – of the demographic variables contained in the so called "Tätigkeitsschlüssel".[3] Thus, we conclude that inconsistencies in the education variable over time are in fact attributable to employers' unreliability and not to institutional changes.

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

We argue that these six educational outcomes can be ranked in increasing order except for the fact that no ranking exists between the second degree VT and the third degree HS. We consider the comprehensive degree HSVT to

overall amount of notifications in order to provide correct aggregate employment statistics (cf. Wermter and Cramer, 1988). Second, there has been a change in the definition of the gross salary subject to social security contributions in 1984 ("Änderungsverordnung zur 2. DEVO"; cf. Bender *et al.*, 1996, and Fitzenberger, 1999, appendix). A third major revision of the notification procedure came into effect in 1999 ("Datenerfassungs- und übermittlungsverordnung"). Henceforth, it has been required for all employers to submit for all their employees subject to social insurance contributions uniform information, which is suited for automated processing.

[3]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 though has not been implemented so far.

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

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 nondecreasing over time because the employees can only attain higher degrees over time, not lose them. Therefore, a decreasing sequence violates this reporting rule and represents evidence for inconsistent reporting behavior. If, instead, employers had to report the educational degree required for the job, a decreasing sequence of education records would not be inconsistent. 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 we weight the spells by their length. 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 of the employee (age, sex, marital status, and nationality), employment status, occupation,[5] industry, number of records the employer gave about this employee, length of the employment spell, and year. The estimation is based on employment 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

[5]We thank Alexandra Spitz for providing a convenient classification for occupations.

part-time workers with less than half the regular hours compared to fulltime 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 for a person across spells are more difficult to deal with since incorrect information is not immediately apparent. Therefore, it is crucial to analyze the sequence of reported education records across spells. If for a person 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 can be the first one overreporting or the second one underreporting or even both can be incorrect. But due to the panel nature of the data, we can identify whether an entire sequence is consistent. In the sample, 81.9% of the persons exhibit consistent sequences of education information while 18.10% do not.

Example 1 shows a person 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 report violates the rule of reporting the highest attained degree. Either, what seems likely in this example, the report of TC itself is wrong, or, in fact, the employee holds the TC at spell 3. Then, all education reports at later spells reporting lower education with ND or VT are wrong because TC is a higher degree. In this example, there exists a second inconsistency. The report of ND at spell 16 is lower than the report of VT at spell 15.

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.[6] 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. However,

[6]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	2	VT	9	yes
16	1	ND	10	yes
17	2	VT	9	yes

Example 1: Person with inconsistently reported education.

they are 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 the employer. 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, the much higher stability of reports by the same employer may also indicate that employers just replicate their previous reports causing serial correlation of reporting errors for reports on the same employee by the same employer.

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%.

When analyzing the probability of missing education reports we find large differences between industries (cf. table 2). To explore further whether there are similar patterns for inconsistent sequences of education reports we look at 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 wrong, we look at characteristics at the first and the second spell: employment status, occupation, industry, spell length and number of reports by the same employer on the employee in consideration. Additionally, we control for age, sex, nationality, and year. The results of the probit regression can be found in table 5. The covariates describing the employment status have the largest coefficients. Working as a trainee (apprentice obtaining a VT) at 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 at 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 stands in sharp contrast with the influence that these variables have on the probability of a missing report. In fact, being foreign or working in the main construction trade or consumer services has a strong positive influence. Section five returns to the question of inconsistent reports, analyzing explicitly the incidence of underreports by comparing the original data and the imputed data.

3 Imputation Procedures

This section develops three imputation procedures to improve the education variable in the IABS. We first discuss the extrapolation of degrees, which is a common feature of all imputation procedures, and then we describe the

three imputation procedures in detail.

3.1 Extrapolation

All imputation procedures developed in this paper are based on the following hypotheses:

(i) after having attained an educational degree, individuals keep their degree,

(ii) the educational degree remains in general almost constant once a person has entered working life and

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

Knowing when a person attained which educational degrees is sufficient to construct the correct sequence of educational degrees for all spells. Therefore, all imputation procedures extrapolate a valid educational degree observed for some spell to all future spells if the future spells report a lower degree or no degree at all until the person attains a higher valid degree or until the last spell if the person does not attain any higher degree. The imputation procedures developed below differ with respect to the heuristic rules applied to determine the valid educational degrees to be extrapolated.

There are three possible types of reporting errors: underreporting education, overreporting education, and not reporting education at all. The existence of missing values is explained by employers who do not spend effort to get to know the education of their employees. They report "information cannot be obtained". The impact of these three types of errors varies substantially when applying an extrapolation rule to the reported data. One has to distinguish between extrapolating the misreported information to subsequent spells and the extrapolation of previously reported information to the misreported spell. Provided one starts with a valid educational degree, extrapolation will correct errors due to underreporting and missing information. However, overreported spells have no chance of being corrected by extrapolating information from previous spells. Furthermore, extrapolating an overreported degree will acerbate the situation. Thus, applying the extrapolation rule to all reported spells on the one hand reduces problems from underreported education or

missing information, but on the other hand, through a ratchet effect, potentially increases problems with overreported education by extrapolating the overreports to later spells. Put differently, a comprehensive application of the extrapolation rule to the highest reported degrees is likely to yield an upward bias by extrapolating overreported degrees.

Imputation procedure 1 (IP1) assumes that there is no overreporting and therefore extrapolates the highest reported degrees. Based on the above discussion, IP1 possibly overstates education through a ratched effect when overreporting occurs. The imputation procedures 2 and 3 (IP2 and IP3) allow for the possibility of overreporting. IP2 and IP3 use heuristic rules to identify valid education information in a conservative way. Only extrapolating valid education information, there is a reduced risk of extrapolating a high degree incorrectly. We argue that, by construction, IP1 tends to overreport education whereas IP2 and IP3 tend to underreport education. Put 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.

The next subsections describe the imputation procedures IP1, IP2, and IP3 in detail. Their implementation comprises four steps. Step 1 defines which spells are accepted as a basis for the extrapolation rule. Education information from not accepted spells is treated as missing and will be imputed by extrapolation from previous spells. Step 2 implements the extrapolation rule from earlier to later spells. Step 3 involves backward extrapolation from later to earlier spells. Step 4 contains further adjustments. The imputation procedures differ only by the acceptance rules used in step 1.

3.2 Imputation Procedure 1 (IP1)

For a given person, IP1 extrapolates every degree which is reported for the first time and is higher than the degrees reported previously. This procedure is justified by the assumption that no overreporting errors occur because the only source of misreports are those employers who report the degree an employee needs for a certain job instead of the highest degree actually attained by the employee. Either such employers just do not check whether

the employee holds a higher degree than needed or they intentionally do not report the higher attained degree, for instance, if the higher degree does not correspond to the lower social status of the job.

For IP1, the extrapolation starts with the first report of a degree. Under the assumption of no overreporting, underreporting of degrees and non-reporting are the only problems. Then, the extrapolation of a (high) degree to spells with missing or lower education information would be a conservative rule that potentially understates true education. In fact, a degree could have already been attained before its first record in the data or an even higher degree as the one reported could have been attained. Since individuals in the data are observed for quite a long time period and people rarely change their education at higher ages, it is, however, likely that the true (or higher) education level is reported eventually. Note that, in a second step, extrapolation also proceeds backwards based on the assumption that people do not change their education at higher ages. However, if there exists a certain probability of overreporting, too, the potential extrapolation of overreports by IP1 may lead to an overstatement of the true education. If there is a remaining bias, we argue that the second effect is likely to dominate, because any plausible sequence of educational records has to be nondecreasing (ratchet effect).

All together, IP1 consists of the following four steps.

Step 1: Extrapolation Rule

IP1 assumes that every first report of a new higher degree in an employment spell is a valid information to be extrapolated subsequently. The level of education in subsequent spells is imputed by extrapolating the most recent valid report. All imputation procedures treat education reports in unemployment spells and other non-employment spells as missing information. The actual extrapolation procedures in steps 2 and 3 will impute education information for the non-employment spells. We do not use information from non-employment spells since it is not reported directly by an employer but repeated from the most recent employment spell. The original data include educational degrees for persons below the age of 18 years which often seem implausible. Therefore, we impute ND for all spells in this age range.

Step 2: Forward Extrapolation

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 one degree is accepted as a valid degree for the first time with the other degree having been accepted before.

We implement the extrapolation rule based on valid degrees in the following order. First, we extrapolate ND to subsequent spells with missing information. Then, we extrapolate VT to subsequent spells with missing information or no degree. Afterwards we extrapolate HS to subsequent spells with missing information or no degree. Now, we impute HSVT if HS is accepted for the first time and VT has been accepted before or vice versa. Finally, we extrapolate HSVT, TC, or UD to subsequent spells with lower or missing information. As a result, degrees are extrapolated starting at the spell of their first acceptance and stopping at the first spell where an even higher degree is accepted.

Step 3: Backward Extrapolation

The forward extrapolation in step 2 leaves the education information missing when spells with missing values precede spells with valid educational information. Since the educational degree of a person is rather time constant, we also extrapolate backwards the first valid educational degree to previous spells with missing information. We do not extrapolate backwards degrees beyond degree specific age limits, because the attainment of a certain degree implies a certain amount 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 backwards UD 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 spells before. Note that the age limits imply that the first spells of young persons can remain with missing information even if these persons at subsequent spells have non-missing information.

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 such an employment status. Therefore, we impute VT for those cases. Subsequently, we also extrapolate the imputed information VT forward and

backwards analogous 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 this is inconsistent with the reporting rule. Hence, in a last step, we impute the highest education information among the parallel spells to these parallel spells.

Example 2 illustrates the implementation of IP1.

SPELL	BILD	Education	IP1	IP1 Education
1	-9	missing	2	VT
2	-9	missing	2	VT
3	-9	missing	2	VT
4	2	VT	2	VT
5	1	ND	2	VT
6	1	ND	2	VT
7	2	VT	2	VT
8	3	HS	4	HSVT
9	3	HS	4	HSVT
10	3	HS	4	HSVT
11	6	UD	6	UD
12	2	VT	6	UD
13	2	VT	6	UD

Example 2: IP1

The forward extrapolation (*Step 2*) extrapolates VT from spell 4 to spells 5 and 6, where the lower education level ND is reported. At spell 8, 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 9 and 10. For spell 11, UD is reported. Even though it is reported only once for this person, IP1 extrapolates UD to spells 12 and 13 because IP1 assumes no overreports occur. Forward extrapolation alone would leave spells 1 to 3 with missing information. Hence (*Step 3*), we extrapolate backwards VT from spell 4 to spells 1 to 3. 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. For example, if there is only missing education information, the person can be left with missing information after the imputation procedures.

3.3 Imputation Procedure 2 (IP2)

IP2 does not rule out the possibility of overreporting. However, without exogenous information, it is not possible to identify spells with overstated education. Instead IP2 uses heuristic rules to identify spells with valid education information and spells with invalid information. Then, only valid education information is used for extrapolation. Information from spells with invalid education is treated as missing and the extrapolation rule imputes information from earlier spells. The critical issue is to identify in a conservative way spells which are likely to be overreported.

IP2 uses the frequency of reporting as an indicator for reliability. Educational degrees which are frequently reported for the same person are likely to be valid. Spells reporting degrees which are very rarely reported for this person are considered incorrect. Concretely, we assume that, if a degree is reported at least three times, then it is reliably reported. We extrapolate this degree from the first time it is reported (if it is higher than the valid reports before). If a degree is reported only once or twice we consider that it is not reliably reported. Invalid information is treated as missing information, hence not extrapolated but instead imputed with information from earlier spells.

When counting specific education reports, we only count employment spells, not non-employment spells. This reflects the fact that only information at employment spells is directly employer reported.

However, under certain circumstances, a low frequency of a specific report arises quite naturally without indicating overreporting. One example involves persons who attain a degree and soon afterwards attain a higher degree. Another example are degrees which are attained shortly before the panel ends. However, this incorrect classification as unreliable spells is by construction limited to two spells. Nevertheless, we try to quantify the problem by implementing two versions of IP2, version IP2A and IP2B. IP2A strictly classifies rarely reported information as invalid. IP2B does so only for persons with

inconsistencies in their reported education. For persons with no inconsistencies, even rare reports can be extrapolated analogously to IP1.

Compared to IP1, IP2A and IP2B yield possibly lower imputed educational reports since not all reported higher degrees are extrapolated. By construction, imputed values for IP2B lie between IP1 and IP2A.

IP2A and IP2B proceed by the same four steps as described above for IP1. The only differences involve step 1:

For extrapolation, IP2A accepts all employment spells as valid when the reported degree is reported in at least three spells. If the total number of employment spells for a person is only four, the minimum frequency for acceptance is reduced to two reports. If there are less than four employment spells, educational information in every employment spell is treated as valid.

IP2B uses the same heuristic rule for extrapolation as IP2A but only for persons with an inconsistent sequence of educational reports in the original data. For persons with a consistent sequence, all employment spells are accepted as a basis for extrapolation.

Both IP2A and IP2B do not accept degrees for persons below the age of 18 but impute ND instead. For young persons below the age of 23 years in vocational training the educational information no degree or only high school degree is accepted even without being reported frequently enough.

Example 3 illustrates the implementation of IP2A.

SPELL	BILD	Education	Employed	Frequency of report	IP2A	IP2A Education
1	1	ND	yes	7	1	ND
2	1	ND	yes		1	ND
3	5	TC	yes	1	1	ND
4	1	ND	yes		1	ND
5	1	ND	yes		1	ND
6	2	VT	yes	3	2	VT
7	2	VT	unemployed		2	VT
8	2	VT	yes		2	VT
9	2	VT	yes		2	VT
10	-9	missing	yes		2	VT
11	1	ND	yes		2	VT
12	1	ND	yes		2	VT
13	1	ND	yes		2	VT

Example 3: IP2A

First, it is determined which degrees are reported at least three times to be valid for extrapolation. ND is reported seven times and hence valid. TC is only reported once. Thus, spell 3 will be treated as a spell with missing information. VT is reported in three employment spells and hence also considered as valid. Actually, it is reported four times in the example, but spell 7 is an unemployment spell just repeating information from spell 6. Hence, spell 7 is treated as a spell with missing information. Now, extrapolation can proceed. ND is extrapolated from spell 2 to spell 3 previously containing the invalid TC report. VT is extrapolated from spell 9 to spell 10 with missing information and spells 11-13 with the lower report of ND. Since the reported information in this example is inconsistent, IP2B would proceed the same way.

3.4 Imputation Procedure 3 (IP3)

Analogously to IP2, IP3 does not rule out the possibility of overreporting. Again, in a first step, a heuristic rule is applied to identify spells with valid education information and spells with invalid information. Then, in subsequent steps, only valid education information is considered for extrapolation.

As an alternative approach to takeing the observed frequency of an educational degree as an indicator of its reliability, we now discriminate between

a series of reports issued by a single employer and reports coming from different employers. The rationale behind this is the hypothesis that 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. If the reporting errors made by individual employers are serially correlated, degrees reported repeatedly by different employers can be considered more reliable than degrees reported repeatedly by the same employer. Tables 3 and 4 might be interpreted this way. Put differently, under this scenario, every change in the educational degree reported for a given person, may reveal us additional information about the reliability of the employer issuing the report and/or the likely true education of the employee.

IP3 is based on an explicit hypothesis about the reporting behavior of the employers. The procedure distinguishes between reliable information from employers who report carefully and unreliable information from employers who do not report carefully. Employers who at most once change the information they report about an employee are classified as reliable. Employers who change the reported degree for a given employee twice or more often are classified as unreliable. This classification is justified by the fact that the educational degree is typically constant for most people during most of their working life. It is not very likely for an employee to attain two (or more) higher degrees while being employed with the same employer.[7]

Furthermore, IP3 tries to explicitly take into account self correction by an employer. We allow for two types of self corrected reporting errors. The first type consists of errors corrected immediately: an employer changes the reported degree only for one spell and immediately afterwards switches back to the degree reported before. In this case, we proceed as if the one time deviation in reporting behavior has never occurred and do not count two but zero changes. Hence, the employer still can be classified as reliable. The second type of self corrected reporting errors concerns reliable employers. If they inconsistently change their report from a higher degree to a lower degree, we interpret this as a valid correction by the employer who actually always intended to report the lower degree. We proceed by assuming that

[7]In our data set, we cannot 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. Thus, we only evaluate the reporting behavior of an employer concerning a specific employee.

the employer always reported the lower degree. If a reliable employer permanently changes to a higher degree we interpret this as the actual attainment of the higher degree.

Now, we start extrapolating a degree when it is reported for the first time by a reliable employer unless the employer did not intend to report this higher degree as a consequence of self correction. In the latter case, we first impute the intended lower degree which becomes the basis for extrapolation.

Again, compared to IP1, IP3 yields potentially lower imputed educational reports since not all reported higher degrees are extrapolated. In this respect, one cannot rank IP3 relative to IP2A or IP2B.

Furthermore, IP3 proceeds by the same four steps as described above for IP1, IP2A, and IP2B. The only differences regard step 1.

For a given individual, IP3 preliminarily accepts all employment spells with non-missing education information which are the first reports of a given employer and, in addition, all employment spells by the same employer whenever the reported degree changes. When a change occurs, the type of reporting error is classified. If there is an immediate self correction, we impute the intended report in the deviating spell. Then, if we count not more than one change in the reported degree, the respective employer is classified as reliable. Otherwise, the employer is classified as unreliable and his reports are set to missing. Next, inconsistent reports by reliable employers are corrected in the spells which were first accepted. As in all other procedures, we impute spells from persons below the age of 18 with ND.

¿From this point onwards, extrapolation proceeds as for the other imputation procedures.

Example 4 illustrates the implementation of IP3.

SPELL	BILD	Education	Employer	Employer reliable	Intended report	IP3	IP3 Education
1	1	ND	1	yes	1	1	ND
2	1	ND	1		1	1	ND
3	2	VT	1		2	2	VT
4	2	VT	2	no		2	VT
5	3	HS	2			2	VT
6	4	HSVT	2			2	VT
7	2	VT	3	yes	2	2	VT
8	4	HSVT	3		2	2	VT
9	2	VT	3		2	2	VT
10	2	VT	3		2	2	VT
11	4	HSVT	3		4	4	HSVT
12	6	UD	4	yes	5	5	TC
13	5	TC	4		5	5	TC
14	5	TC	4		5	5	TC
15	5	TC	4		5	5	TC
16	4	HSVT	5	yes	4	5	TC
17	4	HSVT	5		4	5	TC

Example 4: IP3

Before extrapolation can take place the reliability of the employers has to be determined and self corrected reporting errors have to be detected. Employer 1 changes the reported information once and is hence reliable. Employer 2 changes the reported information twice, thus he is not reliable. His spells (3-6) are treated as spells with missing information. Employer 3 seems to change the reported level of education three times. But we interpret the report of HSVT at spell 8 as an immediate correction of a one time misreport because VT is reported by this employer at spells 7 and 9. Hence, we count only one change and classify this employer as reliable. We conclude that he intended to report VT at spells 7-10 and HSVT at spell 11. Employer 4 changes the reported degree once and is hence reliable. But the report of TC after UD is inconsistent. IP3 assumes this to be a self correction and that employer 4 always intended to report TC. Employer 5 never changes the reported education and is classified as reliable. Now, extrapolation can take place on the basis of the reliable employer's intended reports (i. e. after taking account of the self correction of reporting errors at spell 8 and 12). VT is extrapolated from spell 3 to spells 4-6. TC is extrapolated from spell 15 to spells 16 and 17.

4 Applications

This section compares the corrected education data resulting from the different imputation procedures to the original data. First, we analyze the distribution of the education variable. Second, we look at wage inequality between and within skill groups. Third, we investigate the impact on estimating a wage regression with education as explanatory variable. Furthermore we analyze the incidence of underreports.

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 fulltime 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 do 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 a 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 that 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 to 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 on how the different procedures work can be gained from looking at the conditional imputation probabilities for the different education categories given the reported education. These imputation matrices are reported in tables 8 to 11 based on the raw spells because the procedures are based on unweighted employment spells. The tables are transformation matrices in which the diagonal elements give the probability an original report remains unchanged by the imputation procedure and the off diagonal elements in each row give the probability it is imputed with one of the other education categories. All procedures impute spells containing missing information with ND in about 25% of all cases and with VT in about 50% of all cases. The large values on the diagonals in the tables show that all procedures leave at least 73% of the non-missing reports 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. ND reports are quite likely to be changed, too. 77.4-83.7% of them are unchanged. 15.7-21.3% are imputed with VT. Even if the broad picture looks similar there are differences between the procedures. Only IP1's entries below the diagonal are all close to zero. This reflects the fact that every reported non-missing degree is used for extrapolation under the assumption of no overreporting. Procedures IP2 and IP3, on the contrary, do not extrapolate every degree but some reported degrees classified as unreliable are imputed with lower degrees. This mainly concerns HS which is in 1.3-4.4% of the spells imputed with ND and in 4.5-6.1% with VT. This also concerns HSVT reports. Those reports stand a 3.3-5.7% chance of being imputed with VT.

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 educational degree [TC and UD] (see for instance Fitzenberger, 1999). This makes imputations within these groups irrelevant but tables 8 to 11 show a considerable number of imputations taking 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 12 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 some DEM lower compared to 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 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 13. Imputing the education variable has a notable 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 and 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 change in wage inequality for this group rises considerably from 0.045 to 0.073-0.124. But note the large differences between the procedures in the 1997 estimates, and hence also in 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, however, there are noticeable differences as well.

Table 14 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 measured within group wage inequalities are changed 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 almost not 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 only some measures of wage inequality, especially in the lower part of the wage distribution.

4.3 Mincer-type Earnings Regressions

This subsection investigates the effects of imputing education on estimated wage regressions. Furthermore, we investigate whether and how the measurement error in education is related to wages. We estimate the following Mincer-type earnings equation (Mincer, 1974):

$$\begin{aligned} \log w &= \alpha + \beta_{ND} d_{ND} + \beta_{HS} d_{HS} + \beta_{HSVT} d_{HSVT} \\ &\quad + \beta_{TC} d_{TC} + \beta_{UN} d_{UN} + \beta X + \varepsilon \end{aligned}$$

where $\log w$ is the log daily wage for fulltime employed men in West Germany in 1995. The education variables are dummies for five categories, the most frequent category VT is the omitted category. We also control for age and age squared in X. Since wages are censored from above at the social security threshold, we estimate a Tobit model. The results without further controls are given in table 15. It can be seen that the coefficients for ND, TC, and UN do not differ much between the original data and the different versions of the imputed data. The differences are below 8.7% of the coefficient obtained based on the original data. But the differences are partly significant due to small standard errors.[8] The intercept as a measure for the

[8] We do not estimate the sampling variance of the difference when applying the different imputation procedures. However, if the sample variance of the coefficient is small in all cases then the variance of the difference is small because of the Cauchy-Schwarz-Inequality.

VT log wage also does not seem to differ significantly. But the coefficients for the smaller education categories HS and HSVT change. The difference is largest for HSVT between the original data and the data based on IP1, the coefficients are 0.196 (SE 0.007)[9] and 0.123 (0.005), respectively. IP3 gives results comparable to IP1 and IP2 gives results between IP1 and the original data. The coefficient for HS does not differ significantly due to large standard errors.

We have repeated the estimations with additional controls for being foreign, six occupations and 13 industries (see table 16). The picture remains qualitatively the same. The additional controls reduce the coefficients in absolute value for the log wage differences. Again the coefficients for ND, TC, and UN are quite comparable. The intercept does not differ notably. The coefficients on HS differ but not significantly. The coefficients on HSVT again differ significantly. It is 0.086 (0.006) for the original data and differs most strongly for IP1 with 0.035 (0.005).

Regarding these regressions, the impact of correcting the education variable is fairly small. This is somewhat surprising since in tables 8-11 the composition of the individual education categories differs between the original data and the imputed data. For instance, with probability 21.3%, IP1 classifies a spell with reported education ND as VT. It is not surprising that the imputation procedures do not change the intercepts as estimates for the log wage for VT given the large size of this group and the low rate of change for spells reporting VT initially. The difference is largest for HSVT which is the education category affected most by the imputation procedures.

Finally, we analyze the relationship between the individual wage and the incidence and the type of misreport. This is of some importance since wage estimations in the spirit of Kane, Rouse and Staiger (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 somewhere 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 overreport dummy, which is one if the report in the original data is higher, and an analogous underreport dummy. If measurement error is independent

[9]Here and in the following standard error in parentheses.

of the wage conditional on the 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 17. The coefficient for a missing report varies between -0.097 (0.005) and -0.114 (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 varies across the different imputation procedures. For IP1, it is -0.331 (0.011). This must be due to a very small number of outliers, because IP1 in general does not allow for overreports. For IP2A and IP2B, the coefficients are -0.180 (0.012) and -0.211 (0.017). For IP3, the coefficient on overreported education is insignificant. 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. It seems that measurement error and wage are not conditionally independent given the 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 educa-

tion 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 18. 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 probability 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 only for 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.[10]

5 Conclusions

The education variable in the IAB employment subsample 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 rather 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 on 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 the results of the procedures come close to each other, this should give credibility to our approach. 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

[10]The results for the other imputation procedures are not displayed in the paper but can be made available on request.

skill groups, and we compare wage regressions using the original data and the corrected data.

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 (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 vocational training plus high school. In most dimensions, wage inequality is only slightly affected by the imputation procedures. But 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. In Mincer-type wage regressions, improving the data makes only a small difference, except again for the small category vocational training plus high school. However, wages for three skill groups are typically lower at all degrees when using the imputed data.

References

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

Bender, S., A. Haas, and C. Klose (2000). "IAB employment subsample 1975–1995", *Schmollers Jahrbuch*, 120, 649–662.

Bender, S., A. Bergemann, B. Fitzenberger, M. Lechner, R. Miquel, S. Speckesser, C. Wunsch (2005, forthcoming). "Über die Wirksamkeit von Fortbildungs- und Umschulungsmaßnahmen", *Beiträge zur Arbeitsmarkt- und Berufsforschung*, IAB, Nürnberg.

Card, D. (1999). "The Causal Effect of Education on Earnings", in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics Volume 3*, Amsterdam: Elsevier.

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: Physica Verlag.

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

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

Lewbel, A. (2003). "Estimation of Average Treatment Effects With Misclassification", Working Paper.

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.

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

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

Wermter, W. and U. Cramer (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 Education Variable *BILD* in the Original Data**

Education (abbreviation)[a]	Coded as	Number of spells	Share of spells	Weighted share[b] Empl. men 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

Notes: [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 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/forman	-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
≤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≤30 days	0.0059	(0.0002)**	public administration	0.0025	(0.0004)**
30<spell≤180 days	0.0103	(0.0001)**	agriculture (t-1)	-0.0021	(0.0004)**
spell≤30 days (t-1)	0.0087	(0.0002)**	basic industry (t-1)	-0.0010	(0.0003)**
30<spell≤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/forman	-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 ≤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/f (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 ≤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)*
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	original 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	Original 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: **Imputation Matrix for Procedure IP1**

				Imputed data				
Original data	Missing	ND	VT	HS	HSVT	TC	UD	Total
Missing	19.61	24.77	49.30	0.87	2.50	1.38	1.57	100.00
ND	0.04	77.41	21.29	0.47	0.58	0.12	0.09	100.00
VT	0.03	0.27	95.83	0.00	2.56	0.83	0.47	100.00
HS	0.03	0.64	0.04	76.84	18.91	1.26	2.29	100.00
HSVT	0.02	0.10	0.04	0.01	89.64	5.04	5.14	100.00
TC	0.03	0.04	0.03	0.00	0.00	92.39	7.50	100.00
UD	0.04	0.02	0.01	0.01	0.01	0.00	99.92	100.00
Total	1.90	23.41	63.78	1.07	3.63	2.61	3.60	100.00

Notes: The table contains the conditional probabilities the column information will be imputed given the spell originally contains the row information. Based on all 8,614,058 spells.

Table 9: **Imputation Matrix for Procedure IP2A**

				Imputed data				
Original data	Missing	ND	VT	HS	HSVT	TC	UD	Total
Missing	28.39	24.98	42.45	0.68	1.25	1.01	1.22	100.00
ND	0.53	82.79	16.17	0.26	0.16	0.05	0.04	100.00
VT	0.30	1.53	96.10	0.06	1.25	0.46	0.29	100.00
HS	4.56	4.37	4.53	73.82	10.11	0.91	1.70	100.00
HSVT	0.92	1.88	5.67	1.79	82.88	3.29	3.56	100.00
TC	0.56	0.69	2.61	0.24	0.54	90.48	4.88	100.00
UD	0.42	0.29	1.09	0.25	0.36	0.49	97.10	100.00
Total	3.10	25.68	62.13	1.03	2.47	2.30	3.28	100.00

Notes: The table contains the conditional probabilities the column information will be imputed given the spell originally contains the row information. Based on all 8,614,058 spells.

Table 10: **Imputation Matrix for Procedure IP2B**

				Imputed data				
Original data	Missing	ND	VT	HS	HSVT	TC	UD	Total
Missing	20.43	27.79	46.99	0.84	1.51	1.12	1.33	100.00
ND	0.17	83.66	15.68	0.25	0.16	0.05	0.04	100.00
VT	0.10	0.89	97.00	0.06	1.23	0.45	0.27	100.00
HS	2.10	3.87	4.53	77.12	9.94	0.86	1.58	100.00
HSVT	0.42	1.30	4.46	0.88	86.31	3.18	3.43	100.00
TC	0.29	0.46	2.07	0.16	0.44	91.84	4.74	100.00
UD	0.22	0.19	0.83	0.15	0.30	0.37	97.93	100.00
Total	2.09	25.80	62.89	1.06	2.54	2.32	3.30	100.00

Notes: The table contains the conditional probabilities the column information will be imputed given the spell originally contains the row information. Based on all 8,614,058 spells.

Table 11: **Imputation Matrix for Procedure IP3**

				Imputed data				
Original data	Missing	ND	VT	HS	HSVT	TC	UD	Total
Missing	20.94	24.22	48.93	0.87	2.23	1.30	1.51	100.00
ND	2.53	78.97	17.55	0.39	0.41	0.09	0.06	100.00
VT	0.83	0.78	95.57	0.04	1.76	0.65	0.37	100.00
HS	2.37	1.32	6.09	72.97	14.31	1.08	1.86	100.00
HSVT	2.26	0.56	3.31	0.33	84.87	4.13	4.53	100.00
TC	1.14	0.26	1.74	0.13	0.51	90.21	6.01	100.00
UD	0.51	0.06	0.52	0.09	0.24	0.46	98.11	100.00
Total	3.24	24.09	62.77	1.03	2.99	2.45	3.43	100.00

Notes: The table contains the conditional probabilities the column information will be imputed given the spell originally contains the row information. Based on all 8,614,058 spells.

Table 12: **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
		50	142	140	138	140	141
		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 of ND and HS, M comprises of VT and HSVT; H comprises of TC and UD. The 50th and the 80th wage percentile for H cannot be reported because the wage data is right censored.

Table 13: **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 wages between skill groups at specific wage percentiles based on the wage values from table 12 .

Table 14: **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 wages within skill groups between the respective percentiles based on the wage values from table 12.

Table 15: **Mincer-type Earnings Regression (Tobit) 1**

	Original data		IP1		IP2A		IP2B		IP3	
ND	-0.205	(0.002)	-0.193	(0.003)	-0.196	(0.003)	-0.197	(0.003)	-0.187	(0.003)
HS	0.057	(0.018)	0.007	(0.021)	0.016	(0.019)	0.016	(0.019)	0.035	(0.020)
HSVT	0.196	(0.007)	0.123	(0.005)	0.176	(0.006)	0.174	(0.006)	0.135	(0.006)
TC	0.430	(0.005)	0.414	(0.005)	0.433	(0.005)	0.434	(0.005)	0.424	(0.005)
UD	0.473	(0.005)	0.479	(0.005)	0.496	(0.005)	0.497	(0.005)	0.485	(0.005)
age/10	0.211	(0.005)	0.193	(0.005)	0.192	(0.005)	0.194	(0.005)	0.189	(0.005)
age_sq/100	-0.016	(0.001)	-0.014	(0.001)	-0.014	(0.001)	-0.014	(0.001)	-0.014	(0.001)
intercept	4.554	(0.009)	4.557	(0.009)	4.570	(0.009)	4.562	(0.009)	4.563	(0.009)
lnsigma	-1.144	(0.004)	-1.100	(0.004)	-1.115	(0.004)	-1.109	(0.004)	-1.101	(0.004)
N	145483		157935		156042		157668		156757	
censored	16789		17549		17466		17537		17407	

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. Spells weighted with their length. Robust standard errors clustered at the person level are in parentheses.

Table 16: **Mincer-type Earnings Regression (Tobit) 2**

	Original data		IP1		IP2A		IP2B		IP3	
ND	-0.125	(0.002)	-0.112	(0.003)	-0.122	(0.003)	-0.118	(0.003)	-0.109	(0.003)
HS	-0.018	(0.017)	-0.054	(0.020)	-0.057	(0.018)	-0.052	(0.018)	-0.035	(0.019)
HSVT	0.086	(0.006)	0.035	(0.005)	0.072	(0.006)	0.070	(0.006)	0.042	(0.005)
TC	0.245	(0.005)	0.221	(0.005)	0.240	(0.005)	0.240	(0.005)	0.230	(0.005)
UD	0.320	(0.005)	0.307	(0.005)	0.331	(0.005)	0.330	(0.005)	0.314	(0.005)
age/10	0.174	(0.005)	0.165	(0.005)	0.163	(0.005)	0.165	(0.005)	0.161	(0.005)
age_sq/100	-0.013	(0.001)	-0.012	(0.001)	-0.012	(0.001)	-0.012	(0.001)	-0.011	(0.001)
foreign	-0.064	(0.003)	-0.081	(0.003)	-0.071	(0.003)	-0.075	(0.003)	-0.080	(0.003)
farmer	-0.212	(0.010)	-0.225	(0.010)	-0.222	(0.010)	-0.223	(0.009)	-0.222	(0.010)
service worker	-0.032	(0.006)	-0.056	(0.006)	-0.051	(0.006)	-0.055	(0.006)	-0.056	(0.006)
sales worker	0.171	(0.006)	0.176	(0.006)	0.172	(0.005)	0.172	(0.005)	0.176	(0.006)
clerical worker	0.227	(0.003)	0.244	(0.003)	0.236	(0.003)	0.237	(0.003)	0.244	(0.003)
admin worker	0.281	(0.003)	0.297	(0.003)	0.287	(0.003)	0.289	(0.003)	0.296	(0.003)
agriculture	-0.012	(0.005)	-0.012	(0.005)	-0.012	(0.005)	-0.013	(0.005)	-0.012	(0.005)
basic industry	0.011	(0.003)	0.008	(0.003)	0.010	(0.003)	0.010	(0.003)	0.008	(0.003)
consumer goods	-0.090	(0.003)	-0.093	(0.003)	-0.090	(0.003)	-0.091	(0.003)	-0.093	(0.003)
food industry	-0.112	(0.005)	-0.117	(0.005)	-0.115	(0.005)	-0.116	(0.005)	-0.117	(0.005)
main construction	-0.047	(0.003)	-0.052	(0.003)	-0.050	(0.003)	-0.051	(0.003)	-0.053	(0.003)
constr completion	-0.128	(0.004)	-0.126	(0.004)	-0.126	(0.004)	-0.127	(0.004)	-0.125	(0.004)
trade	-0.179	(0.004)	-0.190	(0.003)	-0.186	(0.003)	-0.188	(0.003)	-0.190	(0.003)
transport & comm	-0.131	(0.004)	-0.152	(0.004)	-0.149	(0.004)	-0.150	(0.004)	-0.153	(0.004)
business services	-0.113	(0.004)	-0.127	(0.004)	-0.126	(0.004)	-0.128	(0.004)	-0.129	(0.004)
consumer services	-0.337	(0.009)	-0.375	(0.009)	-0.365	(0.009)	-0.374	(0.009)	-0.376	(0.009)
education	-0.195	(0.004)	-0.202	(0.004)	-0.202	(0.004)	-0.203	(0.004)	-0.204	(0.004)
public admin	-0.183	(0.004)	-0.186	(0.004)	-0.185	(0.004)	-0.185	(0.004)	-0.188	(0.004)
intercept	4.653	(0.009)	4.659	(0.009)	4.667	(0.009)	4.662	(0.009)	4.665	(0.009)
lnsigma	-1.239	(0.004)	-1.207	(0.004)	-1.216	(0.004)	-1.212	(0.004)	-1.209	(0.004)
N	141,860		153,431		151,769		153,119		152,258	
censored	16,602		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 parenthesis.

Table 17: **Mincer-type Earnings Regression (Tobit) 3**

	IP1		IP2A		IP2B		IP3	
ND	-0.122	(0.003)	-0.120	(0.003)	-0.119	(0.003)	-0.117	(0.003)
HS	-0.029	(0.020)	-0.039	(0.018)	-0.034	(0.018)	-0.017	(0.019)
HSVT	0.091	(0.005)	0.103	(0.006)	0.101	(0.006)	0.086	(0.005)
TC	0.247	(0.005)	0.253	(0.005)	0.252	(0.005)	0.250	(0.005)
UD	0.327	(0.005)	0.340	(0.005)	0.339	(0.005)	0.328	(0.005)
age/10	0.166	(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	-0.122	(0.004)	-0.106	(0.005)	-0.112	(0.004)	-0.120	(0.004)
underreport	-0.113	(0.003)	-0.106	(0.004)	-0.104	(0.004)	-0.116	(0.003)
overreport	-0.331	(0.011)	-0.181	(0.012)	-0.211	(0.017)	0.008	(0.020)
foreign	-0.065	(0.003)	-0.065	(0.003)	-0.067	(0.003)	-0.068	(0.003)
farmer	-0.216	(0.009)	-0.216	(0.009)	-0.217	(0.009)	-0.215	(0.009)
service worker	-0.046	(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	0.228	(0.003)	0.227	(0.003)	0.227	(0.003)	0.230	(0.003)
admin worker	0.281	(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	0.013	(0.003)	0.013	(0.003)	0.013	(0.003)	0.013	(0.003)
consumer goods	-0.088	(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	-0.037	(0.003)	-0.039	(0.003)	-0.038	(0.003)	-0.039	(0.003)
constr completion	-0.124	(0.004)	-0.124	(0.004)	-0.124	(0.004)	-0.123	(0.004)
trade	-0.177	(0.003)	-0.177	(0.003)	-0.177	(0.003)	-0.178	(0.003)
transport & comm	-0.139	(0.004)	-0.138	(0.004)	-0.138	(0.004)	-0.139	(0.004)
business services	-0.116	(0.004)	-0.117	(0.004)	-0.118	(0.004)	-0.119	(0.004)
consumer services	-0.356	(0.009)	-0.350	(0.009)	-0.357	(0.009)	-0.359	(0.009)
education	-0.197	(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	4.673	(0.009)	4.673	(0.009)	4.664	(0.009)	4.664	(0.009)
lnsigma	-1.221	(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 parenthesis. *reportmiss, underreport* and *overreport* are defined in comparison to the original data.

Table 18: **Probit Regression of Underreport compared to IP2A**

Regressors	Marg eff	Robust SE	Regressors	Marg eff	Robust SE
≤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≤30 days	0.016	(0.001)**
50-59 years	-0.005	(0.001)**	30<spell≤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/forman	-0.024	(0.001)**	year 82	-0.006	(0.000)**
home worker	0.042	(0.008)**	year 83	-0.005	(0.000)**
part time ≤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%.