Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

11/2010
EN

Methodological aspects of labour market data

Which factors safeguard employment?

An analysis with misclassified German register data.

Laura Wichert,
Ralf A. Wilke

Bundesagentur für Arbeit

Which factors safeguard employment? An analysis with misclassified German register data.*

Laura Wichert†

Ralf A. Wilke‡

December 2010

*We thank Aderonke Osikominu, an associate editor and two referees for helpful remarks on the paper and the team of the IAB-FDZ for support with the data. Wilke is supported by the Economic and Social Research Council through the *Bounds for Competing Risks Duration Models using Administrative Unemployment Duration Data* (RES-061-25-0059) grant and by the German Research Foundation through the *Statistical Modelling of Errors in Administrative Labour Market Data* grant. This work uses the IAB Employment Subsample (IABS 2004-R04) and the Integrated Employment Biographies V.1 (IEBS-SUF V1) of the Research Data Centre at the Institute of Employment Research (IAB).

†University of Konstanz, Department of Economics, Box D 124, 78457 Konstanz, Germany, E-mail: laura.wichert@uni-konstanz.de

‡University of Nottingham, School of Economics and ZEW Mannheim, E-mail: ralf.wilke@nottingham.ac.uk

Abstract

We analyse the main determinants for job separation with transition to unemployment using individual administrative data from Germany. While the sample size is large and the information in target variables is often highly accurate, non-target variables are subject to considerable measurement error due to a lack of relevance for the data generating process. We show that the high degree of misclassification can even persist after comprehensive logical editing and imputation rules were applied. We find that the measurement error has a sizable effect on our estimation results. Long tenure rather than a higher educational qualification appears to be the key ingredient for a safe job in Germany.

Keywords: unemployment risk, nonclassical measurement error, MC-SIMEX

1 Introduction

We use administrative individual data from Germany to analyze the determinants for job separations with subsequent transition to unemployment. Our analysis aims at contributing to several important questions such as: How do education decisions affect job stability? Are immigrant workers more likely to lose their jobs? What are the most important factors which let the individual transition probability from employment to unemployment shrink or even vanish? Knowledge of these main determinants also contributes to explaining the low labour market dynamics in Germany which has one of the lowest transition rates from employment to unemployment among the OECD countries.

Previous research for Germany and other countries (e.g. Gangl (2003) and Frederiksen (2008)) suggests that attributes associated with individual skills, such as the educational qualification, the wage level, and the labour market experience, have a considerable negative statistical association with the probability of losing a job. Based on monthly household panel survey data from Germany, Gangl (2003) finds evidence for the conditional transition rate to unemployment to more than halve if an individual has Abitur (diploma from German secondary school qualifying for

university admission) or a higher educational qualification rather than not having a completed degree or vocational training. The effect of education is found to be much bigger than the effect of past individual labour market experiences, while a very low wage is associated with a considerably higher risk of unemployment given everything else equal. He finds a higher risk of unemployment (although insignificant) for individuals with an immigration background. As the German household panel survey data are characterised by considerable recall error regarding the labour market experiences of individuals (Jürges, 2007), we perform a similar analysis with administrative data.

Administrative individual data are gradually becoming a prime resource for policy evaluation and empirical labour market research in many countries. This is because the available data sets are large and contain precise information on target variables such as wages, employment periods and the duration and level of employment subsidies and social security transfers. Therefore, it is very attractive for empirical labour market research on the returns to education, wage inequality and the evaluation of labour market programmes, among other things. However, while the administrative data on target variables are generally precise, non-target variables can be subject to considerable measurement errors. In general, administrative data are generated and collected using manifold methods. These include interviews, self-reports and reports from the employer. In some cases, individuals have to present certificates; in others, their reply is entered without any plausibility check. If information is collected solely for statistical purposes, its quality is likely to be lower, since error-checking is labour intensive, and therefore expensive. At worst, this can result in apparent data inconsistencies such as implausible changes in the educational qualification or nationality of an individual over time. For example in Germany, employers report educational qualifications, nationality and job classifications, among other variables, to the public pension insurer for statistical reasons only, yet these variables are irrelevant for the pension entitlements of their employees. Apart from detecting inconsistent information about an individual over time, it is also possible to reveal data inconsistencies if the same variable is available in different administrative sources. While it may only be collected for statistical

reasons in one register, it may be highly relevant information in another source. By validating the lower quality information it is possible to determine the degree of misclassification and the size of the measurement error. Even though there is extensive literature on data quality problems in survey data, only few contributions analyze the quality of administrative data. Several studies compare survey and administrative data to determine misclassification. However, these studies often assume that the administrative data are correct and use them as validation information for the survey data. For example, see Benitez-Silva et al. (2004) for self reported disability status. Kapteyn and Ypma (2007) compare information on earnings in US administrative and survey data. By focusing on wage data, they can assume that the administrative information is generally reliable. Johansson and Skedinger (2009) doubt that the disability information in administrative data is always reliable and find evidence that disability status is misreported in Swedish administrative data.

There is a broad literature on different general methodologies to deal with data problems. While statistical research has often focused on classical measurement error (for a summary see, for example, Cameron and Trivedi, 2005, chapter 26) and regression techniques with incomplete data (Schafer, 1997), here we face an error structure that violates the statistical regularity conditions for classical measurement error. Since we have ordered and non-ordered discrete or binary variables rather than continuous variables, there are natural restrictions on the sign of the measurement error that make it non-classical. While multiple imputation methods (see for example Little and Rubin (1987), Schafer (1997)) primarily focus on the elimination of missing values, there are also methods for editing and imputing data (see for example Fellegi and Holt (1976), Manzari (2004)) which use logical rules or information in neighboring observations to eliminate inconsistencies and missing values. In context of German administrative data, both methods have been applied to different variables. Büttner and Rässler (2008) apply multiple imputation methods to impute missing values due to top coding in the wage variable of the German employment records. Fitzenberger et al. (2006) observe many inconsistencies and implausible changes in the educational qualification in the same data and suggest several editing and imputation corrections closely related to the logic-

driven Fellegi-Holt methodology. Their approach is interesting because it eliminates many apparent inconsistencies. We will apply their rules to our data and we suggest a similar approach for the nationality variable. Moreover, by making use of our derived misclassification information for the education and citizenship, we apply a misclassification SIMEX (MC-SIMEX, Küchenhoff et al., 2006) for the estimation of a nonlinear regression model with misclassified discrete variables. Our program code for the data corrections and our MATLAB implementation of the MC-SIMEX will be made available to the user community of these data by the research data centre of the German Federal Employment Agency (IAB-FDZ, fdz.iab.de).

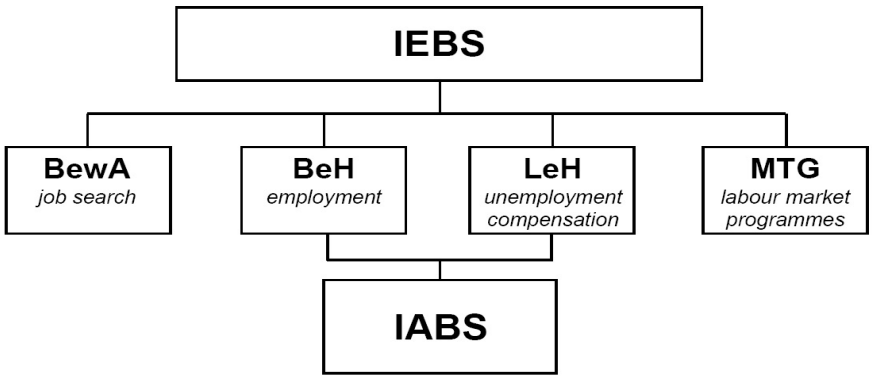
The paper is structured as follows. Section two reviews and introduces the editing rules for the education and citizenship information in German administrative employment records. Section three uses validation data from other administrative sources for a misclassification analysis. In section four, we present the estimation results of our application to unemployment risk. The last section summarizes and defines future research needs.

2 Data and Editing Rules

Since register data are comprised of highly sensitive information, they are often not easily accessible for independent researchers and the user group is therefore in most cases restricted to government contractors or national research institutes. The IAB-FDZ has facilitated access for a wider international user community by offering standardized data products as scientific use files, such as the IAB Employment Sample (IABS) and the Integrated Employment Biographies (IEBS). The IABS contains daily employment records (Beschäftigtenhistorie, BeH) for a 2% random sample of the German workforce subject to social security contributions for the period 1975-2004. In addition to the employment periods, the BeH contains information on the individual (such as gender, age, wage, educational qualification, nationality and job title) and the employing firm (such as the business sector and the location of the firm). These employment spells are linked with daily claim periods for unemployment compensation from the German Federal Employment Agency (Leistung-

shistorie, LeH). For more information on the IABS, see Hamann et al. (2004) or Drews (2008). The IEBS contains the same sources of information as the IABS in the period 1990-2004 but with less or higher aggregated information on the firm and individual level. As an advantage, for the years after 1999 it is linked to the job seekers register (Bewerberangebotsdatei, BewA) and the register of training measures (Massnahmeteilnehmer-Gesamtdatenbank, MTG). See Figure 1 for an illustration of the sources of the two scientific data sets. The IEBS is a 2.2% random sample of the joint population of the four administrative registers. For more information on the IEBS see Zimmermann et al. (2007). While information on education and nationality in the BewA is actually used in the job search process of the unemployed it is collected for statistical reasons only in the BeH. Therefore, the IEBS contains an additional and more reliable source of information that will be used to assess the reliability of the information in the IABS. The IABS is more commonly used in empirical research because it has an easier data structure and covers a longer time period. Moreover, it contains more information related to employment, firms and the region. The IEBS is predominantly used for the evaluation of active labour market programmes. Our empirical analysis uses the IABS as it allows us to work with a richer set of variables.

Figure 1: Sources of German administrative labour market data.



We apply editing and imputation rules for the educational qualification and the nationality in the employment records to eliminate data inconsistencies. We restrict

our analysis to the information in the BeH, as it is a main data source for the IABS and it is the only informative source for education and citizenship in the scientific use file version of the IABS. Since our variables of interest are non-target variables we expect them to contain a considerable amount of measurement error. The literature about editing and imputing discusses several approaches to deal with measurement errors. Manzari (2004) reviews methods for data editing and imputing and applies them to population census data. In her paper, she combines two methods: the Fellegi-Holt methodology (Fellegi and Holt (1976)) and the nearest neighbour imputation methodology (Bankier et al. (1997) and Bankier (1999)). While the first method is logic-driven by applying logical editing rules about one individual to detect inconsistencies, the latter is data-driven and uses information from other individuals (called 'donors') to correct the data. In the present analysis, we apply Fitzenberger et al.'s (2006) correction method for the education variable and we introduce an editing rule for the citizenship variable that identifies individuals with an immigration background. Both imputation procedures are closely related to the logic-driven Fellegi-Holt methodology since they only use within-person information. This method has been proven to perform well in cases of random errors while the nearest neighbor method is more appropriate for systematic errors (Manzari (2004)). Even though we find that there is a tendency to understate the educational level in the data, we assume that the errors in the education and the nation variable can be considered as being random and therefore not deterministic.

Fitzenberger et al. (2006) suggest four imputation procedures for the education variable in the IABS. Based on the idea that an individual's educational level cannot decrease over the life cycle they develop rules to detect inconsistencies in the education variable over time. They introduce four different imputation procedures which differ in the requirements for the educational history to be used to overwrite the inconsistent information in subsequent spells. The authors claim that it is impossible to say which procedure is the best among the four but any of their imputations shall lead to improvements in data quality. We compare the imputed data based on their weakest and strictest rules with validation data (see Section 3). As results were similar but suggest that their imputation procedure 1 (IP1) leads to a bigger

Table 1: Cross tabulation of IP1 (imputed) vs. uncorrected education in the BeH, 20,960,096 spells.

	Education						
IP1	Missing	ND	VT	HS	HSVT	TC	UD
Missing	14.32	.02	.01	.06	.01	.01	.01
ND	24.79	75.12	.27	.68	.10	.05	.01
VT	50.06	23.01	94.51	.05	.03	.01	.01
HS	2.31	.69	.00	73.35	.01	.00	.01
HSVT	4.48	.84	3.46	21.50	87.96	.01	.00
TD	1.83	.18	1.08	1.50	5.77	90.09	.00
UD	2.21	.14	.67	2.86	6.12	9.83	99.96
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Abbreviations: ND: no degree, VT: completed vocational training, HS: high school degree (Abitur), HSVT: high school degree and complted vocational training, TC: technical college degree, UD: university degree.

reduction in measurement error, we only report the results for IP1 in what follows. IP1 tends to impute higher educational levels than the other three suggested rules. Table 1 presents a cross tabulation of IP1 and the original education. It is apparent that the imputation procedure changes from less than 1% (UD) to up to 25% (ND, HS) of the values of the education variable. Almost a quarter of the "No degree" cases are changed to "Vocational training". More than 85% of the "Missings" are eliminated by the imputation procedure, being replaced with "No degree" or "Vocational training" in almost 75% of the cases.

When developing a logical editing rule for the nationality it is important to note that implausible changes of nationality may not necessarily point to data inconsistencies. Because many immigrants have the German nationality in addition to their inherited nationality it is difficult to disentangle misreporting from the event of having multiple nationalities. For this reason our editing rule aims at identify-

ing individuals with an immigration background instead of trying to recover the citizenship in each spell. According to our definition an individual has an "immigration background" if it has more than one non-German nationality spell. If the individual is observed just once, this spell is sufficient. The variable has a missing value for individuals without any information on nationality in the data. Note that the following results are robust with respect to the number of required non-German spells. When cross tabulating the immigration background against the nationality in the BeH we find that all diagonal elements are greater than 0.98. Although our editing rule induces relatively few changes in the data, it is important to note that a large share of spells from individuals with immigration background are recorded as German in the original data. Indeed, our data editing rule is relevant as it increases the number of "non German" spells by about 20%, from 1.78m to 2.13m spells. It will become apparent later that this has a crucial effect on our empirical results.

3 Misclassification Analysis

In this section, we analyse the measurement error in the education and nationality information and we assess the quality of the data correction rules. We determine misclassification with the help of the IEBS by comparing information in the BeH with information in the BewA. If the educational qualification or the nationality in the BeH do not match the information in the BewA, we define this as misclassification. We use for our analysis only spells starting in 1999 or later because BewA information is not systematically available in earlier years. Since information in the BewA is a target variable and is collected for non-statistical reasons, it is considered to be of higher quality than the information in the employment records. Some research on data quality confirms this view (Bender et al., 2005). To confirm this assessment we repeat the editing and imputation analysis of the previous section for the BewA and indeed, we find a considerably lower share of inconsistent observations than in the BeH (6% versus 20% in case of education). Although this does not suggest that our validation data are free of error, there is strong evidence for them being far less erroneous. However, further systematic research is required to

check the validity of the validation data by, for example, using information from other linked administrate sources or survey data if they were available.

Our approach to validating BeH information is based on information in BewA spells if these overlap with BeH spells or if other spells follow promptly. When we choose only those BeH spells which overlap with BewA spells as the validation sample, we are left with 651,261 spells, or about 10.5% of all BeH spells in the period 1999-2004. As the event of having overlapping spells may be rather selective, we also allow for a gap of up to two weeks gap between BeH and BewA spells. In this case, we are left with about 1.2m spells, or about 20% of all BeH spells in the period 1999-2004. As the following misclassification results are very similar for the two samples, we only report them for overlapping spells. As we are interested in misclassification of information in the IABS, we make two modifications to the IEBS to make information in the BeH spells comparable. This includes setting the nationality information to "Missing" for all individuals who have one employment record in Eastern Germany and constructing a comparable educational qualification variable for the BewA. See Appendix A1 for more details.

Tables 2 to 7 contain the misclassification matrices for the variables of interest. The main diagonal elements reflect the share of observations which match in the two variables. It is apparent from Tables 2 and 3 that both education variables (original and IP1) are highly misclassified as many diagonal elements are below 0.5. The tables also suggest that IP1 has reduced the amount of misclassification as diagonal elements tend to be higher.

For further analysis we group the education variable in four categories: "Missing", "No degree", "VT" (Vocational training or any kind of school degree) and "Higher Education" (technical college or university degree). This is done because Tables 2 and 3 suggest that VT, HS or HSVT are often coded as one of these other categories. The same is true for TD and UD. As the educational levels within these two groups of categories are similar anyway, we pool them to further improve the precision of the data. For this reason we obtain the grouped categories "VT" (VT, HS, HSVT) and "Higher education" (TD, UD). Indeed, the diagonal elements of the misclassification matrices for the education variable increase due to the grouping (see

Tables 4 and 5).

For the nationality, Table 6 suggests that only in 72% of the cases the information on non German nationality matches with the BewA information, while it differs in about 27% of the cases. This provides evidence that the measurement error in the foreign citizenship information is greater than commonly believed. In contrast the immigration concept captures the non German information considerably better. Table 7 suggests that the immigration status in BeH and BewA coincides in more than 95% of the spells.

Although, we detect a large amount of misclassification in the data, it is important to note that our results may not hold for the entire German population. This is because we have only employees with recent unemployment experiences in our sample and therefore dropping all employees without unemployment experiences. We check whether this affects our results by investigating in two directions: first, we check whether the more extensive validation sample, which is twice as large as the sample of overlapping spells, produces similar results. In the case of the nationality variable and the imputed education variable, deviations between the misclassification probabilities are very small and less than 1% points in all cases. In the case of the original education variable, the differences are also rather small, but for few values they reach 5% points. Although, our results are robust we cannot repeat this analysis for employees without any unemployment experience. Second, we check whether the descriptive statistics for the sample change if we use our validation samples instead of all BeH spells. We find that they are similar for most variables in the data. The few larger deviations are in accordance that our validation samples consisting only of employees who become unemployed once.

In order to evaluate the quality of our editing and imputation strategies, we follow the guidelines given by Chambers (2006), who presents imputation performance measures for categorical variables used in the EUREDIT project. In particular, we use a measure for the degree of misclassification in the data by computing a weighted share of misclassified observations in the data, with zero being the optimal value (no misclassification). When we compute the measure for the original data we find that both rules improve the data quality. It decreases from 33% to 21% in the case of the

Table 2: Misclassification matrix for education (uncorrected).

Education				BewA			
BeH	Missing	ND	VT	HS	HSVT	TC	UD
Missing	48.65	43.67	30.80	43.43	29.97	25.97	26.26
ND	16.66	32.61	10.64	19.48	7.24	3.86	3.78
VT	28.50	22.76	56.70	18.25	39.37	23.36	17.64
HS	.30	.34	.25	10.19	3.14	2.49	2.97
HSVT	1.63	.35	.86	4.19	11.14	7.39	5.44
TD	2.24	.13	.55	1.83	4.76	22.52	6.24
UD	2.02	.14	.21	2.63	4.37	14.41	37.67
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Abbreviations: ND: no degree, VT: completed vocational training, HS: high school degree (Abitur), HSVT: high school degree and complted vocational training, TC: technical college degree, UD: university degree.

Table 3: Misclassification matrix for IP1 (imputed).

IP1				BewA			
BeH	Missing	ND	VT	HS	HSVT	TC	UD
Missing	9.96	5.01	1.62	7.73	1.94	2.99	3.30
ND	23.29	39.93	7.45	19.92	5.75	2.35	2.60
VT	54.13	51.79	84.80	27.53	38.67	14.88	11.52
HS	.17	.63	.14	14.87	3.68	2.45	3.08
HSVT	4.00	1.90	3.69	20.10	29.35	10.94	7.63
TD	4.13	.38	1.61	4.10	10.30	36.15	7.44
UD	4.33	.36	.68	5.75	10.30	30.25	64.43
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Abbreviations: ND: no degree, VT: completed vocational training, HS: high school degree (Abitur), HSVT: high school degree and complted vocational training, TC: technical college degree, UD: university degree.

Table 4: Misclassification matrix for the grouped education (uncorrected).

Grouped education	BewA			
BeH	Missing	No degree	VT	Higher Educ.
Missing	48.65	43.67	31.02	26.16
No degree	16.66	32.61	10.55	3.81
VT	30.43	23.46	56.88	28.35
Higher Education	4.26	.27	1.55	41.68
Total	100.00	100.00	100.00	100.00

Table 5: Misclassification matrix for grouped and imputed education (IP1).

Grouped IP1	BewA			
BeH	Missing	No degree	VT	Higher Educ.
Missing	9.96	5.01	1.79	3.20
No degree	23.29	39.93	7.59	2.52
VT	58.30	54.31	86.60	24.16
Higher Education	8.45	.75	4.01	70.12
Total	100.00	100.00	100.00	100.00

Table 6: Misclassification matrix for nation (uncorrected).

Nation	BewA		
	Missing	German	non German
Missing	92.70	.03	.15
German	6.29	98.65	27.48
non German	1.01	1.31	72.37
Total	100.00	100.00	100.00

Table 7: Misclassification matrix for immigration background (imputed).

Immigration background	BewA		
BeH	Missing	German	Migration
Missing	97.59	.00	.01
German	1.99	96.28	3.16
Migration	0.42	3.71	96.83
Total	100.00	100.00	100.00

education variable and from 3% to 2% for the nationality/immigration background.

Due to its enormous size, administrative data can be used to perform statistical analysis of some smaller groups such as young people or selected geographic areas. Since misclassification can be more or less pronounced in certain population segments, it is therefore of vital interest to analyse the relationship between the probability of misclassification and other observable variables such as worker and firm characteristics. We perform this analysis by estimating Logit regressions with the dependent variable equals one if an observation is misclassified and equal to zero otherwise. Since the suggested immigration concept has a rather low misclassification probability ($< 5\%$) we only present results for the imputed education variable IP1. Table 8 presents the estimated marginal effects on the probability of misclassification of the grouped and imputed education variable evaluated at the sample means of the other regressors. Although it is apparent that the event of misclassification is related to different variables, it is surprisingly difficult to find a systematic pattern of misclassification determinants that is valid for all models. The predicted probability of misclassification of having no educational qualification is 61% which is similar to the average value (60.07%, Table 5). Being young or a non German decreases this probability by 31% and 12% respectively. While being employed in the construction sector or in trade increases this probability by 9%. These figures suggest that there are certain subgroups in the data with considerably different degrees of misclassification. The predicted probability of misclassification of the grouped category VT is 12%. Being employed in East Germany decreases this probability by 9%, while being young increases this probability by 17%. The predicted probability of misclassification of the grouped category higher education is 29%. Being employed in East Germany decreases this probability by 17%, while being young or being employed in mining increases this probability by 54% and 44%, respectively. These figures provide evidence that the probability of misclassification in the education variable strongly varies across population segments and it can exceed levels of 80% and more. In these cases it will be almost impossible to obtain reliable results with these data. We also include the actual length of the BeH spell as a covariate to analyze whether information in shorter spells is more likely

to be erroneous than in longer spells, since firms may already anticipate the short duration and devote less care in completing the records. This hypothesis is partly supported by the data. While such a pattern is not present for the nationality, there is some evidence for it in the case of the education variable but only for the "No Degree" category where longer spells have a lower misclassification probability than shorter spells. When comparing the average length of BeH spells in our two validation samples, we observe that it is very similar and about 168 days, while it is on average 237 days for all BeH spells in the same period. This suggests that the average misclassification probability in the case of no degree may be considerably lower for an average BeH spell than reported in our tables. Since this deviation is driven by individuals with long employment and no unemployment periods, we have no validation data at hand to investigate this issue further.

Without reporting the results, we also find a positive correlation between misclassification of the education and the nationality variable. This suggests that the reliability of information is likely to vary across reporting firms or individuals. A more detailed analysis would require, however, the availability of a firm identifier.

Table 8: Marginal effects (at the sample mean of other variables) of a Logit regression for the determinants of misclassification in the grouped and imputed education (IP1).

	Dependent variable: Misclassification of ...			
	Education missing	No Degree	VT	Higher Educ.
variable	ME (SE)	ME (SE)	ME (SE)	ME (SE)
female	.0144 (.0009)	.0257 (.0026)	-.0136 (.0011)	.0067 (.0057)
aged <25	.0468 (.0014)	-.3058 (.0032)	.1726 (.0021)	.5367 (.0243)
aged >55	-.0022 (.0016)	-.0166 (.0041)	.0348 (.0018)	-.0530 (.0084)
non German (orig.)	.0062 (.0014)	-.1159 (.0030)	.0569 (.0019)	.1288 (.0103)
part time	-.0431 (.0017)	.0048 (.0028)	-.0039 (.0011)	-.0118 (.0063)
high income	.0014 (.0035)	.0831 (.0086)	.0875 (.0033)	-.0802 (.0073)
low income	-.0573 (.0012)	-.0607 (.0035)	.0373 (.0013)	.1237 (.0069)
business sector, ref: others				
agriculture	.0373 (.0031)	.0517 (.0069)	-.0423 (.0023)	-.0012 (.0237)
mining	-.0939 (.0249)	.0029 (.0258)	.0110 (.0104)	.4417 (.0684)
manufacturing	.0387 (.0013)	-.0391 (.0039)	-.0292 (.0013)	.0507 (.0100)
construction	.0351 (.0016)	.0895 (.0044)	-.0578 (.0013)	.0673 (.0159)
trade	.0273 (.0012)	.0847 (.0036)	-.0354 (.0012)	.1281 (.0105)
gastronomy	.0071 (.0019)	.0510 (.0039)	-.0172 (.0017)	.1321 (.0209)
minor jobs	-.0276 (.0052)	.0793 (.0094)	.0032 (.0039)	.0791 (.0239)
eastern Germany	.0319 (.0016)	.2569 (.0029)	-.0953 (.0019)	-.1730 (.0097)
length of BeH spell, ref: 2-9 months				
up to one month	.0349 (.0015)	.0001 (.0035)	-.0099 (.0014)	.0911 (.0089)
more than 9 months	-.0179 (.0011)	-.0359 (.0029)	.0084 (.0012)	.0110 (.0062)
predicted probability	.9193	.6079	.1205	.2879
Log. likelihood	-3,078.06	-112,256.13	-161,248.65	-17,065.56
Number of observations	10,102	178,857	432,548	29,754

4 Application

In this section we empirically analyze how the educational qualification or the nationality affect the probability of losing a job. Our sample is extracted from the IABS. We reorganize the employment spells in these data into a monthly panel of employees. We estimate a Logit model for unemployment risk, where the dependent variable is 1 if the employee incurs a job loss in the current period and becomes unemployed while it is 0 otherwise. A job loss is defined as observing the beginning of an unemployment compensation claim spell within one month after the end of the employment spell. We do not perform a cross section analysis at one point of time because unemployment inflows have important seasonality patterns. We do not use a panel with a higher frequency (e.g. weekly) because there are almost no cases with two job losses within one month and due to most independent variables being constant within a month. Still by having up to 12 observations per year for each individual, the size of the panel data is intractable for statistical analysis. We therefore restrict it for our analysis to the period 1999-2002. This leaves us with about 20m observations generated by more than 580K employees. Table 11 in the Appendix presents a complete list of variables in the model and the summary statistics of the sample. As we observe the employment history of the individuals since the early 1990s (if not even longer) we construct several variables such as tenure, labour market experience and past unemployment experiences. Moreover, we include information on the current job such as wage. Our model contains more than 50 covariates which also include individual characteristics of the worker, the employing firm, and calendar time. Despite having a panel structure we estimate the model with pooled Logit. In order to have a causal interpretation of model coefficients, it is required that covariates are not correlated with the error term. As this is difficult to test, we consider the estimated model coefficients as statistical relationships between the covariates and the probability of making a transition to unemployment given everything else equal in the model. We do not apply fixed effect estimators to allow for some correlation between the time constant part of the error and the covariates as our key variables (education and nationality) do not vary over time in most cases. The application of differencing techniques such as the logit fixed effects estimator

does not yield meaningful results in this case. We compute heteroscedasticity robust standard errors for clustered data (Williams, 2000).

In addition to the estimated Logit coefficients, we report the relative marginal effect (RME). To compute the RME of a variable j we first calculate the marginal effect on the transition probability in response to a change in variable j at the mean of all other independent variables. This marginal effect is then divided by the predicted transition probability of the reference individual to obtain the RME. RME= 0 therefore corresponds to having no effect at all while RME= 1 suggests that the change in variable j is estimated to double the unemployment risk. We report the RME rather than the marginal effect as the level of the latter depends on the longitudinal unit of the data, while the RME is invariant (for more details see Dlugosz et al., 2009). We perform a sensitivity type analysis by estimating the same model with original and corrected variables to identify the effect of the data corrections on the estimated model coefficients:

- A: original data
- B: corrected education, immigration background.

Table 9 presents the estimated coefficients for the key variables together with their RMEs. By comparing the estimates for the uncorrected variables and the edited variables, we observe large changes. We find evidence that the magnitude of the education effect drops by about a half if we use the imputed education information rather than the original education. According to the results based on the original education data, having no degree increases the probability of losing a job and becoming unemployed compared to the same individual with vocational training by almost one fifth. This number halves to 9% if we use the imputed education variable instead. Higher education decreases the probability of entering unemployment, but the RME of higher education is only -11% for the imputed education variable compared to -20% for the original variable. This is again a drop by one half. For the nationality, the results suggest that non-German individuals have a 3% lower probability of losing their job compared to Germans. The effect changes its sign to 5% if we use the immigration background concept. Missing information on na-

Table 9: Results of Logit regressions

	Model A				Model B		
variable	coeff.	(SE)	RME		coeff.	(SE)	RME
Grouped education, ref: VT				*Grouped and Imputed Education (IP1)*, ref: VT			
no degree	.1725	(.0086)	.1876***	no degree	.0831	(.0097)	.0863***
higher educ.	-.2285	(.0186)	-.2037***	higher educ.	-.1124	(.0152)	-.1060***
missing	.1752	(.0091)	.1908***	missing	-.3960	(.0318)	-.3263***
Nation, ref: German				*Immigration background*, ref: none			
non German	-.0306	(.0109)	-.0300***	Immigration	.0494	(.0099)	.0505***
missing	.1738	(.0130)	.1891***	missing	.1775	(.0131)	.1935***

Note: fully robust standard errors (heteroscedasticity, serial correlation).
***, **, *: marginal effect significant at the 1, 5 and 10% level, respectively

tionality shows the strongest relative effect and increases the unemployment risk by about 19% in both models. As this information is missing for all individuals with at least one employment record in east Germany, it suggests that unemployment risk in East-Germany is considerably higher than in West-Germany.

It is likely that estimated coefficients for the corrected education variable are still biased due the presence of considerable non-classical measurement error. For this reason, we have also estimated a misclassification Logit regression by applying the MC-SIMEX method (Küchenhoff et al., 2006). The results are indicative for the estimated coefficients being biased due to the remaining misclassification but due to the large sample size we were not able to obtain inference statistics. See Appendix 2A for a brief outline of this method and a presentation of first results.

The RMEs of the remaining variables based on the Logit estimator are given in Table 10. These results do not differ substantially between the two models, therefore, we only present the RME's for Model B. We do not find important gender differences in unemployment risk. Age shows a strong effect, older individuals aged

55 or more have a 85% higher probability of loosing their jobs than individuals aged between 25 and 50. This could be due to age discrimination or due to the fact that older workers often use unemployment benefits as a convenient exit route out of regular employment to old age pensions. Among the individual background variables, past unemployment has the strongest effect. If an individual has been unemployed before, his risk of reentering unemployment increases by 136%, more than doubling. According to the descriptive statistics in Table 11 in the Appendix, our sample consists of 37.97% observations of individuals who have been unemployed before. This illustrates the prominent role of past unemployment as the main predictor of entering unemployment. Jobs with low income (defined by having a wage in the bottom quantile of the population distribution of daily wages in west or east Germany, respectively) are also rather unsafe as such individuals face a 83% higher risk of unemployment. Interestingly, the sample correlation between past unemployment and low wage is rather low, although positive. Part time workers, who are mainly female, have a much lower probability of making a transition into unemployment. In our sample, many observations with a part time job are associated with a low wage. This suggest that the high unemployment risk of low wage jobs only applies to male full time workers with a low daily wage. This is likely related to a high wage replacement rate in case of unemployment for this group.

To disentangle the effects of labor market experience and tenure, we construct an experience variable which is total labour market experience net of tenure in the present job, which gives us additional experience. Comparing the results for tenure and experience shows that both have a positive effect on job security, both increasing with the number of years. However, the effect of tenure is much larger. Individuals with more than four years of tenure have an unemployment risk that is 79-88% lower than individuals with no tenure, which corresponds to a predicted probability of almost zero. An equivalent amount of additional experience only lowers the risk by 15-37%. In our sample, almost 50% of the observations are generated by individuals with four or more years of tenure. This large share of individuals with extremely low unemployment risk explains why the overall mean predicted monthly probability is just 0.31%. This is in line with the results of Elsby et al. (2009) who show that

Table 10: Results of the Logit regression - relative marginal effects for Model B

variable	RME		variable	RME	
female	-.0286	***	*month*, ref: June		
aged <25	.0571	***	January	.2928	***
aged 51-55	.2504	***	February	-.1619	***
aged >55	.8139	***	March	-.0362	***
low income	.8703	***	April	-.3333	***
past unemployment	1.3954	***	May	-.3674	***
previously recalled	.6759	***	July	-.0681	***
seasonal job	.3388	***	August	-.2189	***
white collar	-.2684	***	September	-.0967	***
in vocational training	-.4843	***	October	-.1091	***
parttime	-.4740	***	November	-.0746	***
tenure, ref: < 7 months			December	.9477	***
7 - 12 months	.0293	***	*business sector*, ref: agriculture		
13 - 24 months	-.4329	***	goods production	-.0628	***
2 - 3 years	-.5673	***	manufacturing	-.2860	***
4 - 7 years	-.7880	***	steel & car industries	-.2706	***
8 - 14 years	-.8659	***	consumer goods	.0590	**
> 14 years	-.8674	***	drink and tobacco	-.0335	
additional experience[1], ref: < 7 months			construction	.6126	***
7 - 12 months	.0754	***	finishing	.2801	***
13 - 24 months	.0724	***	wholesale	-.0056	
2 - 3 years	-.0452	***	retail	.0032	
4 - 7 years	-.1570	***	traffic	-.1794	***
8 - 14 years	-.2363	***	private services	-.0871	***
> 14 years	-.3610	***	home services	.1415	***
year, ref: 2001			health services	-.1528	***
1999	-.0405	***	public firms/organisations	-.0577	**
2000	-.0921	***	public administration	-.2649	***
2002	.2307	***			
predicted probability	0.0031				
Log. likelihood	-706,558.37				

[1] *additional experience= total experience - tenure*

***, **, *: corresponding marginal effects significant at the 1, 5 and 10% level, respectively

Germany is among the OECD countries with the lowest unemployment inflow rate. As our evidence has a descriptive nature, we cannot distinguish between two possible explanations: first, the strong effect of long tenure may be due to long tenured jobs having a very high level of employment protection in Germany; second, long tenure is a proxy for the high ability of a worker or for firm specific human capital.

We find strong seasonal unemployment patterns, with far fewer separations in April and May and far more in December and January. The spike in the winter separations is, to some extent, due to firms' planned capacity reductions during the winter period, seasonal employment and many work contracts end et at the end of the calendar year. When comparing business sectors, we find evidence that between 1999-2002, the safest jobs were in manufacturing and in public administration, while the construction and finishing works are characterised by a considerably higher separation rate.

When we compare all these effects, it becomes evident that the effect of education on unemployment risk is rather small compared to other individual factors, especially if we use the imputed data. The main indicator for a safe job is long tenure rather than high education. This is in contrast to previous evidence based on survey data (Gangl, 2003) and for other labour markets with higher dynamics such as Denmark, where the educational qualification appears to be far much more important (Frederiksen, 2008). We do not find evidence for discrimination of females and only weak discrimination evidence for individuals with immigration background.

5 Summary and Remarks

We analyze the determinants for job separation with transition to unemployment using German register data, taking into account that non-target variables in the data contain a considerable amount of measurement error. We adapt existing editing and imputation methodologies for the education variable and suggest an additional editing rule for the nationality variable. We use information from an accompanying administrative register to compute misclassification probabilities for the education and the nationality variables and to show that the editing and imputation rules

indeed reduce the amount of measurement error. We provide evidence that the degree of misclassification strongly varies across data segments. Depending on the target group of the analysis, the data may be too erroneous for obtaining even only roughly reliable empirical results.

We perform a sensitivity type analysis to determine whether estimated coefficients change after the imputation, confirming that the correction rules have a strong effect on empirical results. In particular, we observe that the effect of education halves in magnitude when using the imputed data instead of the original data. The effect of not being German changes its sign. Our results therefore suggest that standard results for classical measurement error do not hold for nonlinear models with non-classical measurement error, because there would be no change in the sign of the estimated coefficients and their magnitude would increase after editing and imputing the data. Our findings demonstrate that measurement error in register data can lead to misleading conclusions about the effect of education or foreign nationality on individual labour market outcomes even if the data are large and partly precise.

While individual labour market outcomes are strongly associated with individual skills, our application suggests that it is mainly the length of tenure that eliminates the unemployment risk in Germany. Although a higher educational qualification is related to a safer job, its role seems to be far less important than commonly thought and suggested by previous evidence based on household survey data. By international standards, Germany has a low transition rate from employment to unemployment. Our results suggest that this is mainly due to a large share of the working population with very long tenure. These employees often stay decades with the same employer. Whether this is due to a better employer-employee fit than in other countries, more corporate responsibility of firms in Germany or just a result of the strong dismissal protection cannot be answered by this analysis.

References

Bankier, M. (1999), Experience with the new imputation methodology used in the 1996 Canadian census with extension for future censuses, *Work Session on Statistical Data Editing*, Rome.

Bankier, M., Fillion, J.M., Luc, M., and Nadeau, C. (1997), Imputing numeric and qualitative variables simultaneously, in Statistical Data Editing, vol. 2, Methods and Techniques: Proc. Conf. European Statisticians, Statistical Standards and Studies, 30–38, Geneva, United Nations.

Bender, S., Biewen, M., Fitzenberger, B., Lechner, M., Miquel, R., Osikominu, A., Waller, M., Wunsch, C. (2005) Die Beschäftigungswirkung der FbW-Maßnahmen 2000-2002 auf individueller Ebene: Eine Evaluation auf Basis der prozessproduzierten Daten des IAB - Zwischenbericht Oktober 2005. Goethe University Frankfurt and SIAW St. Gallen.

Benitez-Silva, H., Buchinsky, M., Chan, H.M., Cheidvasser, S. and Rust, J. (2004) How large is the bias in self-reported disability? *Journal of Applied Econometrics*, 19, 649–670.

Büttner, T. and Rässler, S. (2008) Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity. *IAB Discussion Paper, 44/2008*, IAB, Nürnberg.

Cameron, A.C. and Trivedi, P.K. (2005) Microeconometrics. Cambridge University Press, Cambridge.

Chambers, R. (2006), Evaluation Criteria for Editing and Imputation in EUREDIT, in: Statistical Data Editing, Vol. 3: Impact on Data Quality. United Nations Statistical Commission and United Nations Economic Commission for Europe, New York and Geneva.

Cook, J. and Stefanski, L. (1994), Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

Dlugosz, S., Stephan, G. and Wilke, R.A. (2009) Fixing the leak: unemployment incidence before and after the 2006 reform of unemployment benefits in Germany, *ZEW Discussion Paper No. 09-079*. ZEW, Mannheim.

Drews, N. (2008), Das Regionalfile der IAB-Beschäftigtenstichprobe 1975-2004, *FDZ Methodenreport No. 02/2008*, IAB Nürnberg.

Elsby, M., Hobijn, B. and Sahin, A. (2009), Unemployment Dynmanics in the OECD, *NBER Working Paper No. 14617*.

Fellegi, I.P. and Holt, D. (1976), A systematic approach to edit and imputation, *Journal of the American Statistical Association*, 71, 17-35.

Fitzenberger, B., Osikominu, A., and Völter, R. (2006) Imputation Rules to Improve the Education Variable in the IAB Employment Subsample. *Schmollers Jahrbuch*, 126, 405–436.

Frederiksen, A. (2008) Gender Differences in Job Separation Rates and Employment Stability: New Evidence from Employer-Employee Data. *Labour Economics*, 15, 915–937.

Gangl, M. (2003) Unemployment Dynamics in the United States and West Germany. Physica-Verlag, Heidelberg New York.

Hamann, S., G. Krug, M. Köhler, W. Ludwig-Mayerhofer, and A. Hacket (2004), Die IAB-Regionalstichprobe 1975-2001: IABS-R01, *ZA-Information*, 55, 36–42.

Johansson, P. and Skedinger, P. (2009) Misreporting in register data on disability status: evidence from the Swedish Public Employment Service. *Empirical Economics*, 37 (2), 411–434.

Jürges, H. (2007). Unemployment, life satisfaction and retrospective error, *Journal of the Royal Statistical Society: Series A*, 170, 43-61.

Kapteyn, A. and Ypma, J.Y. (2007) Measurement Error and misclassification. A comparison of survey and register data. *Journal of Labor Economics*, 25, 513–551.

Küchenhoff, H., Mwalili, S.M. and Lesaffre, E. (2006) A general Method for Dealing with Misclassification in Regression: The Misclassification SIMEX, *Biometrics*, 62, 85–96.

Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J. Wiley and Sons, New York.

Manzari, A. (2004), Combining Editing and Imputation Methods: An Experimental Application on Population Census Data, *Journal of the Royal Statistical Society. Series A*, 167 (2), 295–307.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data, *Biometrics*, 56, 645-646.

Zimmermann, R., Kaimer, S. and Oberschachtsiek, D. (2007), Dokumentation der Integrierten Erwerbsbiographien (IEBS-SUF V1), *FDZ Methodenreport No. 01/2007*, IAB Nürnberg.

Appendix

Table 11: Descriptive statistics

variable	mean	variable	mean
gender, ref: male		*calendar time*	
female	.4279	*month*, ref: June	
age, ref: 26-50		January	.0822
aged <25	.1450	February	.0821
aged 51-55	.0912	March	.0826
aged >55	.0906	April	.0829
		May	.0831
employment history		July	.0833
past unemployment	.3795	August	.0840
previously recalled	.1036	September	.0846
tenure, ref: < 7 months		October	.0844
7 - 12 months	.0914	November	.0842
13 - 24 months	.1311	December	.0833
2 - 3 years	.1576	*year*, ref: 2001	
4 - 7 years	.1656	1999	.2444
8 - 14 years	.1670	2000	.2506
> 14 years	.1410	2002	.2512
additional experience[1], ref: < 7 months			
7 - 12 months	.0299	*business sector*, ref: agriculture	
13 - 24 months	.0547	goods production	.0574
2 - 3 years	.1224	manufacturing	.0910
4 - 7 years	.1905	steel & car industries	.0787
8 - 14 years	.2081	consumer goods	.0528
> 14 years	.1082	drink and tobacco	.0271
		construction	.0351
current employment		finishing	.0285
low income	.3543	wholesale	.0592
seasonal job	.1507	retail	.0822
white collar	.4050	traffic	.0516
in vocational training	.0616	private services	.1450
part-time	.1605	home services	.0485
		health services	.1084
		public firms/organisations	.0562
		public administration	.0572
original education, ref: vocational training		*IP1*, ref: vocational training	
no degree	.1791	no degree	.1397
high education	.0823	high education	.1106
missing	.1065	missing	.0132
original nation, ref: German		*immigration background*, ref: German	
non German	.0815	immigration	.1090
missing	.0347	missing	.0343
number of observations	20,659,889		
number of individuals	582,698		

[1] *additional experience= total experience - tenure*

A1: Construction of a validation variable for the educational level The BewA contains two different variables describing the educational level of a person: the schooling level (*schbild*) as well as the professional level (*bild*). In order to compare the imputed values based on the LeH- and the BeH-spells with the information given in the BewA, we first have to recode the two variables of the latter to a corresponding single variable. For this purpose, we chose two rules: first, the "strict version" requires valid information in both sources, and second, the "weak version" relies more on the information in the bild-variable, and accepts missings in the schbild-variable. We think that the latter version is also justifiable, because the employer is not so much interested in the schooling level, but more in the highest completed degree, which is either a vocational training or an university or technical college degree. Since there is no big difference between the two variables (only in about 0.05% of the spells), we only use the "weak version" for the following analysis. Table 12 illustrates the construction of the new validation variable for education.

Table 12: Recoding scheme of the education variable ("weak version") in the BewA for the education validation variable (EDU_val)

BewA			EDU_val
schbild		*bild*	
No school degree or at most Mittlere Reife[1] or missing	AND	No vocational training degree	ND
No school degree or at most Mittlere Reife[1] or missing	AND	Vocational training degree but no technical college nor university degree	VT
Fachabitur[2] or Abitur[3]	AND	No vocational training degree	HS
Fachabitur[2] or Abitur[3]	AND	Vocational training degree but no technical college nor university degree	HSVT
Fachabitur[2] or Abitur[3] or missing	AND	Technical college degree	TD
Fachabitur[2] or Abitur[3] or missing	AND	University degree	UD
Any value	AND	Missing	Missing

[1] minimum 10 years of schooling (general certificate of secondary education)

[2] minimum 12 years of schooling (vocational diploma)

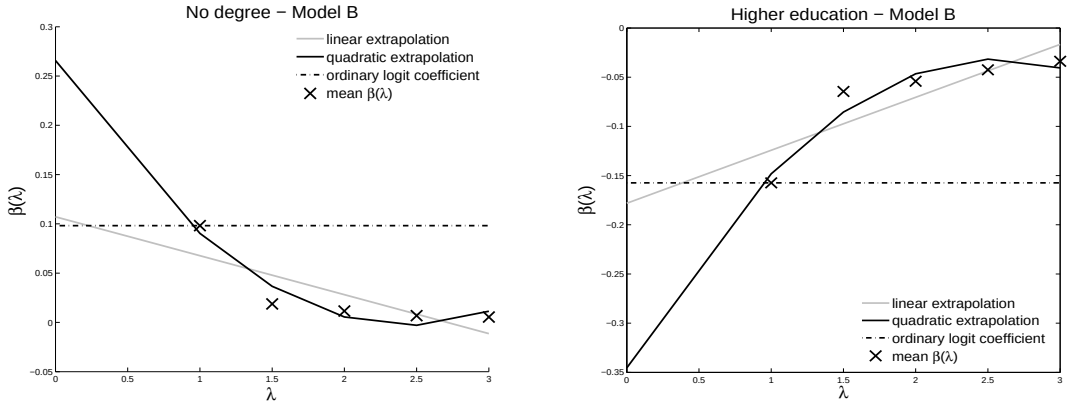
[3] minimum 13 years of schooling (general qualification for university admission)

A2: The MC-SIMEX The MC-SIMEX (Küchenhoff et al. (2006)) can be applied to (non)-linear regression models in presence of measurement error in discrete variables. It is a modification of the SIMEX algorithm for additive measurement error (Cook and Stefanski (1994)). The following informal presentation of the MC-SIMEX uses the imputed education variable in Model B for a better illustration of the method. Table 13 contains the misclassification matrix for IP1 in our application and Figure 2 contains a graphical illustration of the estimation procedure.

Table 13: Misclassification matrix for IP1 in Model B, 462,560 spells.

Education	Validation data		
IP1	No degree	VT	HE
No degree	42.81	7.67	2.62
VT	56.42	88.17	24.89
HE	0.77	4.16	72.49
Total	100.00	100.00	100.00

Figure 2: Fitted extrapolants and ordinary logit estimator.



The algorithm works in two steps: in the first step it simulates new data for the erroneous variables by increasing the size of the measurement error in the data. If we

consider the observed variable as having one "degree" of misclassification ($\lambda = 1$), the simulated data has a higher degree of misclassification. The simulation is done for several degrees of misclassification, i.e. for $\lambda = 1.5, 2, 2.5, 3$. The model is then re-estimated by using the more erroneous variable at each step (i.e. for each λ), while all other variables are unchanged. Then, new data for the erroneous variable is generated by further increasing the degree of misclassification and the model is again re-estimated, and so on. These simulation and estimation steps are repeated 200 times for each degree of misclassification, i.e. for each λ-step. Then, the mean of all the estimated coefficients is kept for each degree of misclassification (mean $\beta(\lambda)$, denoted by "X").

In the second step, the estimator in the case of no measurement error is obtained by an extrapolation from the simulation results in presence of misclassification. Sticking to the notation that the observed misclassified variable contains one degree of misclassification, the case of no misclassification can be seen as a zero degree of misclassification, i.e. $\lambda = 0$. Accordingly, we fit an OLS curve through the mean of the coefficient estimates of each simulation step (i.e. through mean $\beta(\lambda)$). The estimated value of the coefficient in absence of misclassification is obtained by an extrapolation of the fitted curve to the value of zero misclassification. There are several functional forms thinkable for the extrapolation function. Küchenhoff et al. (2006) suggest to use the linear and the quadratic extrapolation function, which are both presented in Figure 2. Based on first inspection, the quadratic extrapolant seems to have the best fit in all cases and is therefore chosen. In the case of the imputed education, the ordinary logit estimate for "no degree" is about 0.1 and -0.15 for "higher education". The coefficients obtained by using the quadratic extrapolant are then 0.26 and -0.35, respectively.

Since the MC-SIMEX is very computer intensive, we were not able to obtain results for the whole sample but we used a 30% random sample instead. Even for this smaller sample with about 6m observations, standard errors are not available. Therefore, our results are only indicative for further considerable changes in the estimated coefficients.

Imprint

FDZ-Methodenreport 11/2010

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Herrlinger

Technical production
Dagmar Herrlinger

Download
http://doku.iab.de/fdz/reporte/2010/MR_11-10-EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Ralf A. Wilke
University of Nottingham,
School of Economics
Email: ralf.wilke@nottingham.ac.uk