Kruppe, Thomas; Matthes, Britta; Unger, Stefanie

Working Paper

Effectiveness of data correction rules in process-produced data: The case of educational attainment

IAB-Discussion Paper, No. 15/2014

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Kruppe, Thomas; Matthes, Britta; Unger, Stefanie (2014) : Effectiveness of data correction rules in process-produced data: The case of educational attainment, IAB-Discussion Paper, No. 15/2014, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg

This Version is available at:
https://hdl.handle.net/10419/103072

Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

IAB-Discussion Paper 15/2014

Articles on labour market issues

Effectiveness of data correction rules in process-produced data

The case of educational attainment

Thomas Kruppe,
Britta Matthes
Stefanie Unger

ISSN 2195-2663

Effectiveness of data correction rules in process-produced data

The case of educational attainment

Thomas Kruppe, Britta Matthes, Stefanie Unger (IAB)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

Abstract 3
Zusammenfassung 3
1 Introduction 4
2 Institutional background and theoretical considerations 5
3 Data and methods 11
3.1 Data 11
3.2 Suitability of the educational information in ALWA for assessing data correctness 12
3.3 Comparison of educational information 13
4 Structure of false and missing notifications 14
5 Correction rules 17
5.1 Using the panel structure of within-person employment data 17
5.2 Using the panel structure of within-person employment data and assumptions about the firm's reliability of notification 19
5.3 Using the panel structure of within-person employment as well as job search data 20
5.4 Effectiveness of correction rules: descriptive evidence 20
5.5 Effectiveness: Multivariate Evidence 23
6 Conclusion 28
References 29

Abstract

The use of process-produced data plays a large and growing role in empirical labor market research. To address data problems, previous research have developed deductive correction rules that make use of within-person information. We test data reliability and the effectiveness of different correction rules for information about educational degrees as reported in German register data. Therefore we use the unique dataset ALWA-ADIAB, which combines interview data and process-produced data from exactly the same individuals. This approach enables us to assess how effective the existing correction rules are and whether they manage to eliminate structural biases. In sum, we can state that simple editing rules based on logic assumptions are suitable for improving the quality of process-produced data, but they are not able to correct for structural biases.

Zusammenfassung

Prozessproduzierte Daten spielen eine große und wachsende Rolle in der empirischen Arbeitsmarktforschung. Um Datenprobleme zu verringern, wurden in der bisherigen Forschung deduktive Regeln zur Korrektur entwickelt, die übergreifende Informationen zu einer Person nutzen. Wir testen die Datenreliabilität und die Effektivität verschiedener solcher Korrekturregeln anhand der in deutschen Registerdaten erfassten Bildungsinformation. Dazu nutzen wir den einmaligen Datensatzes ALWA-ADIAB, in dem Interviewdaten und prozessproduzierte Daten von ein und derselben Person kombiniert sind. Aufgrund der hohen Realibilität der Interviewdaten messen wir daran sowohl die unkorrigierten als auch die korrigierten Registerdaten. Dieses Vorgehen erlaubt es uns zum einen, die Effektivität der Korrekturregeln zu bewerten. Zum anderen können wir prüfen, ob durch die Anwendung der Korrekturen strukturelle Verzerrungen behoben werden. Im Ergebnis zeigt sich, dass einfache, auf logischen Annahmen beruhende Datenaufbereitungen zwar geeignet sind, die Qualität der prozessproduzierten Daten zu verbessern, strukturelle Verzerrungen aber verbleiben.

JEL classification: C8, I2

Keywords: measurement error, data quality, process-produced data, register data, interview data, reliability, data correction, imputation, missing values, structural bias, educational degrees, labor market research

1 Introduction

One of the most striking features of recent decades has been the rising importance of process-produced data as a source for empirical labor market research (c.f. Baur 2011; Hochfellner et al. 2012; Kröger et al. 2011; Røed and Raaum 2003; Seysen 2009; Wagner 2012). This rise has occurred because the available datasets are large and contain precise information on variables such as wages, social security transfers, starting and ending dates of employment and unemployment periods. Indeed, some of the variables in process-produced data are of extraordinary reliability. However, some of them are error-prone because they contain divergent information, inconsistent individual sequences or even missing values. Therefore, the reliability of some process-produced data has been doubted (Scioch and Oberschachtsiek 2009:242).

Furthermore, the reliability of process-produced data depends on the purpose for which this specific information is gathered. In this article, we will provide examples of problems with data on educational achievement in the process-produced dataset SIAB[1], provided by the Research Data Centre of the Federal Employment Agency at the Institute for Employment Research. In short, in this dataset some information on educational achievement of an individual originates from employers' notification to the German social security system. Whereas reporting this information is only required for statistical purposes, it is largely unverified in the notification process. Moreover, misreporting has no consequences concerning obligations or claims for social security - neither for the employer nor for the employee. Additionally, other information on educational attainment originates from administrative procedures for registering job searchers at the German Federal Employment Agency. Because the information on educational achievement is very important for placing people in the labor market, it is presumed to be very reliable. To address divergent information from the two sources, including inconsistent individual sequences or even missing values, previous researchers have developed deductive correction rules, making use of assumptions about the sequential order of states during an employment career and about the credibility of the informant (Drews 2006; Fitzenberger et al. 2006; Kruppe 2006; Wichert and Wilke 2010). All methods that are applied take more or less plausible assumptions as a starting point. Nevertheless, we know neither how effective these correction rules are nor whether any of them have the ability to correct for potential structural biases. Until now, there was no way to assess the reliability of the corrected education variables arising from the different

[1] We have used German administrative data made available by the German Federal Employment Agency (Bundesagentur für Arbeit) and the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung). This data is confidential and cannot be made available as public use data set. The data and programs used for the paper can be accessed by interested researchers at the research data center of the Federal Employment Agency (http://fdz.iab.de). Please contact: mailto:iab.fdz@iab.de for further details.

correction rules. However, by utilizing a new combined dataset called the ALWA-ADIAB, we are now able to perform such an assessment. The first source of the ALWA-ADIAB, the survey dataset ALWA, gathered the individual educational and employment history of respondents in every detail. The second source of the ALWA-ADIAB is the process-produced dataset SIAB. As discussed above, it contains information about educational attainment submitted by employers as well as recorded through the official registration of job-seekers. Because the ALWA-ADIAB is a dataset linked to both sources on an individual basis, the information on educational achievement from ALWA can be compared with that from the process-produced SIAB for each respondent. As a result, we are able to directly measure the success of different correction rules for process-produced data in terms of improvements to the reliability of this information.

The article is structured as follows: First (section 2), we look at the institutional background, i.e., the notification processes for the German social security system and the internal procedures of the German Federal Employment Agency. Based on this information, we identify reliability problems concerning information on educational achievement in process-produced data. Section 3 describes the data we used in detail. We also explain why the information on educational achievement collected in the survey dataset ALWA is highly reliable and therefore suitable for assessing whether the corresponding information in the process-produced data is correct. This section ends with methodological remarks on how we will analyze the quality of information on educational attainment and test the effectiveness of different correction rules. Section 4 outlines, in a descriptive manner, the ideas behind various correction rules and their approach to and success in improving the reliability of process-produced data, as well as applies multivariate logistic regressions to the dataset to assess whether the correction rules manage to eliminate structural biases. In section 5, we discuss the impact of the results on future research.

2 Institutional background and theoretical considerations

There is extensive literature on data quality problems in survey data (for an overview see Lyberg et al. 2012; Schnell 2011), but few studies analyze the quality of administrative data. Several studies have made use of them to validate survey data (Benitez-Silva et al. 2004; Gottschalk and Huynh 2010; Johansson and Skedinger 2009; Kapteyn and Ypma 2007; Kreuter et al. 2010) based on the assumption that these process-produced data are correct or at least highly reliable. However, some studies have detected inconsistencies or implausible sequences in administrative data (Bernhard et al. 2006; Bollinger and David 2005; Fitzenberger et al. 2006; Huber and Schmucker 2009; Jaenichen et al. 2005; Scioch 2010), and some studies have focused on how to address missing data (Büttner and Rässler 2008). In addition to these more technical considerations, it has been argued that such inconsistencies exist because of differences between the underlying definitions of

particular measures in the process-produced data (Davies and Fisher 2008; Johnson and Moore 2008). Kruppe (2009) showed that the data quality in process-produced data is strongly influenced by the institutional setting in which the data are collected, e.g., the underlying measurement concepts. Using two different definitions of unemployment, he showed the effects of these definitions and their implementation on data quality regarding the calculation of unemployment duration in Germany.

To reach a deeper understanding of the reliability problems of process-produced data, we first have to look at the origin of these data. Let us start with the notification procedure for data on such employment which is subjected to social insurance contributions (see Figure 1).

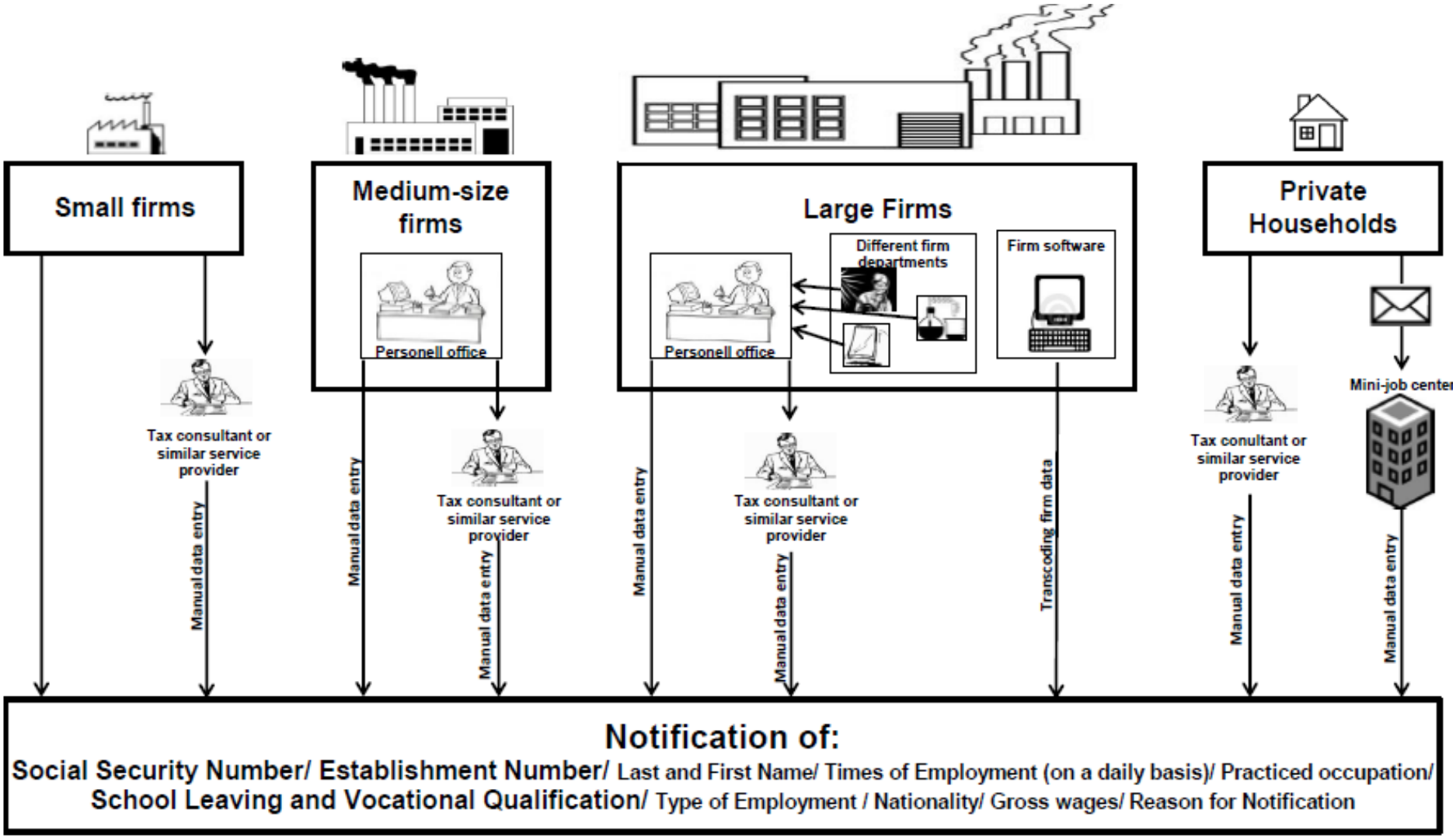
In Germany, all employers have to report specific details of employment for every employee covered by the social security system (notification requirement) to the social security system. As observed in Figure 1, there are various notification procedures. Depending on firm size or type of employer (firm/private household), a specific notification procedure is preferred: Whereas tax consultants or similar service providers are relevant for all types of notification procedures, medium-sized and large firms often are equipped with personnel offices that directly notify the social security system of changes. Private households might send reports about their staff to a specific mini-job center on their own, or they may consult with a service provider who then implements the notification procedures.

These notifications must be made at the beginning and end of each employment period. Additionally, the notifications are mandatory for temporary employment interruptions, changes in health insurance or even annually, if no change occurs. The main reason for gathering these data is for making wage-based calculations of the individual's level of contributions to and resulting claims for social security. Additional information reported by the employers, such as the educational achievement of the employee, is a non-target variable. As a consequence, these variables are not checked for errors and there are no incentives for accurate reporting. Employers know that they report some information only for statistical purposes and that misreporting neither has consequences (in terms of obligations or claims) for the employer nor does it pose a financial risk or come with other disadvantages to the employee. As a result, such variables in process-produced employment data are assumed to be less reliable.

In contrast, unemployment insurance data are produced through the internal processes of the Federal Employment Agency and consist of receipts for unemployment benefits and data from registered job searches and participation in labor market programs. These programs are meant to improve participants' opportunities for finding a new job. The process of data collection is computer-aided. Information about starting and ending dates of unemployment and participation in

labor market programs is very accurate because the amount of unemployment benefits or subsistence payments provided depends on these data. Likewise, information on the educational achievement of registered job seekers should be very reliable as it used to place the unemployed into the labor market. In some cases, job seekers even have to present their educational certificates. Therefore, educational achievement is a target variable here, and we consider these data to be of higher reliability than the information on educational achievement in employment records.

Figure 1
Notification procedure for data on employment



Source: own illustration.

However, in process-produced data, we are not only confronted with different levels of reliability but also with periods without any information: Some groups are excluded from the collective forms of the social security system. There is a gap in the data if people become civil servants, freelancers or self-employed, or if they move into the education system or leave the labor force.

In sum, we differentiate between five ideal-typical reliability problems that can arise in process-produced data. Although these problems occur in all types of process-produced data, we exemplify them with information on educational achievement (cf. Figure 2).

The five reliability problems are as follows:

1. *Inconsistencies in parallel episodes from the same data source:*

If a person has two jobs in two different firms, it can happen that, for example, one firm reports that the employee holds a vocational degree, whereas the other firm reports that the same employee holds a university degree.

2. *Inconsistencies in episodes from different data sources:*

It may happen that a person is employed but also registered as a job seeker at the same time. Therefore, two differing specifications about the same information may be reported simultaneously.

3. *Missing information*

An employer may report that a certain person is employed but may fail in reporting additional information on educational attainment.

4. *No information*

If someone becomes a civil servant, freelancer, or self-employed, or moves into the education system or leaves the labor force, there is a gap in the data.

5. *Inconsistencies over time*

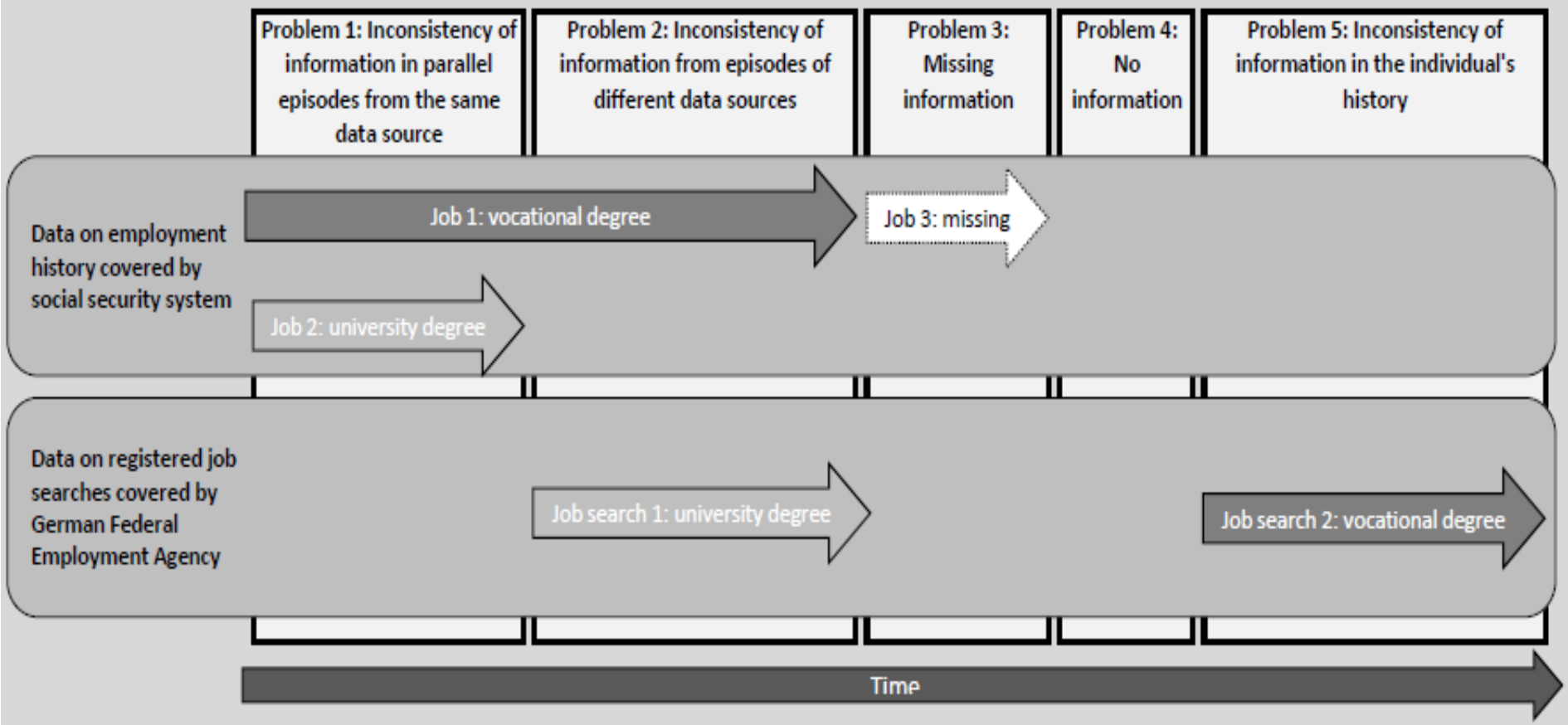
Looking at an individual's history, there may be inconsistencies over time, e.g., during several periods of unemployment (or employment), and a higher educational degree may be reported at an earlier point in time than what has most recently been reported.

Thus, educational attainment in process-produced data is error-prone. Because we do not know how important the reliability problem is, we must

a) look at the number of correct matches and mismatches and the amount of missing information on educational attainment (section 4) and,
b) check whether, by using rather simple correction rules, inconsistencies and missing information in process-produced data can be removed in most cases and whether possible structural biases are correctable (section 5).

In the following section, we will first describe in detail the data we used.

Figure 2
Reliability problems of process-produced data



Source: own illustration.

3 Data and methods

3.1 Data

In this article, we utilize a new unique dataset called the "ALWA Survey Data linked to Administrative Data of the IAB" (ALWA-ADIAB), which combines survey information from the ALWA study ("Arbeiten und Lernen im Wandel" - Working and Learning in a Changing World) with the "Sample of Integrated Labor Market Biographies" (SIAB), which has been generated from process-produced social security records, as well as data on the work establishments of the respondents from the "Betriebshistorikpanel" - Establishment History Panel (BHP) (see Figure 3).

The ALWA data include a wide range of longitudinal variables, especially detailed longitudinal information about entire education and employment histories, as well as many cross-sectional variables regarding, for instance, gender, birth date, religion and parental background of the respondents that have been gathered from more than 10,400 computerized telephone interviews. The sample was drawn from the universe of German residents belonging to the birth cohorts of 1956-1988, regardless of their nationality (Antoni et al. 2010; Kleinert et al. 2011; Matthes and Trahms 2010).

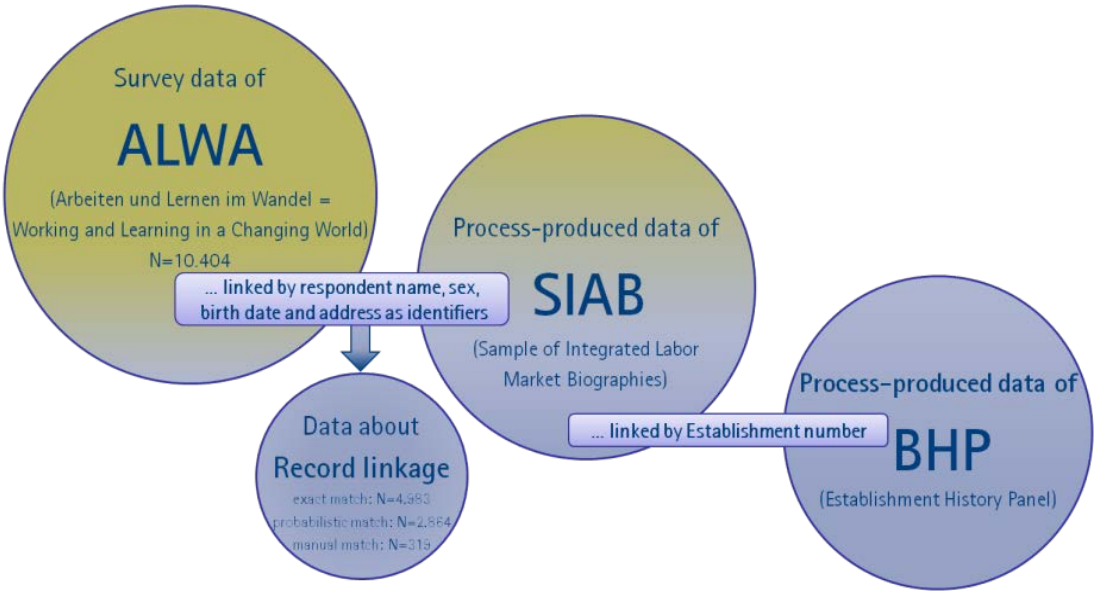
The administrative part of the ALWA-ADIAB consists of comprehensive data on the individual employment histories of ALWA respondents how it can be found in the Sample of Integrated Labor Market Biographies (SIAB) dataset.[2] The individual employment histories consist of information on employment and unemployment benefits from as far back as 1975 and job searches starting in 2000 (Dorner et al. 2010). These data were merged with establishment data (BHP), thus enriching them substantially. The establishment data provide information on the respondents' employment firms on a yearly basis, including industrial classification codes, dates of birth and death of the establishment, number of employees, median daily wage, the balance of genders, distribution of ages and qualification structures of the establishment, and so on (Hethey-Maier and Seth 2010). In this paper, we make use of this information to analyze the effect of firm size on notification behavior and the quality of information provided.

However, not all ALWA participants are included in the linked dataset. This is mainly because every respondent had to provide an informed consent to approve the linkage of the survey data with the administrative data. Additionally, he or she had to be identified on a one-to-one basis in the administrative records of the Federal

[2] Not every ALWA respondent is contained in the SIAB, as the latter is a 2 percent random sample of the Integrated Employment Biographies (IEB). Therefore, we have included all of the available IEB data for the ALWA respondents who consented to the linkage of the survey data with the process-produced data in the same way as was done by drawing the SIAB from the IEB. Thus, the documentation of the SIAB can be used to report on this part of the data structure and the variables contained.

Employment Agency. The ALWA sample was drawn from the registers of the residents' registration offices in German municipalities. Therefore, the link of an individual's information from both datasets is based on the respondent's name, sex, birth date and address. If no one-to-one identification was possible, fault-tolerant record linkage techniques were applied to maximize the number of linked respondents by making allowances for spelling mistakes. As a result, information from 8,166 ALWA respondents could be linked to the administrative data (Antoni and Seth 2012).

Figure 3
ALWA-ADIAB



Source: own illustration.

3.2 Suitability of the educational information in ALWA for assessing data correctness

To assess whether information on the educational attainment of a certain person in the process-produced dataset was correct or not, we compared their level of educational attainment registered in the SIAB with that of the same person in the ALWA survey. It appears to us that the information on educational achievement in ALWA is highly reliable for three reasons. First, ALWA is a retrospective survey that asks for educational history in great detail - not only for achieved educational degrees. It is very unlikely that degrees can be forgotten, as degrees were collected during the report about a specific period of educational life. For example, the survey of school history begins with questions about the first school attended and asks for related information (such as the beginning and ending dates of school attendance, the achievement of a completion certificate, which certificate could have been achieved). Then, it asks whether respondents attended any additional schools – including adult education courses - that led to further school degrees. If the respondent attended another school, again, detailed information was collected. This

procedure continued until all schools attended were surveyed. After that, the respondent was asked about further qualifications that he or she may have achieved through an external examination or through completing vocational training. If the respondent had done so, information on the name of the degree and the date of achieving it was collected.

Second, in the ALWA survey, respondents reported about their lives - not their achievements. The detailed questioning about certain periods of the respondents' lives prevented mistakes and misunderstandings. Third, to ensure high reliability of the ALWA data, several cross-checks were implemented. In addition to commonly used checks of value ranges, a data check-and-revision-module was implemented in order to detect and resolve inconsistencies immediately. In collaboration with the respondents, the interviewers were encouraged to ask for additional information during the interview if something was unclear (Drasch and Matthes 2011).

3.3 Comparison of educational information

Based on the arguments presented in section 3.2, we can assume that the data on educational attainment are correct in the ALWA survey. Thus, we can compare this information to the educational information in the process-produced data. However, the educational variables are not identical in both datasets. The main difference is that the educational information in the ALWA is gathered as a history of educational events consisting of schooling episodes, episodes of vocational preparation as well as vocational, professional and academic education. In contrast, in the process-produced SIAB dataset, educational information is stored as panel information from employment and job search episodes.

Taking this difference into account, we calculate the highest educational attainment for every respondent from the ALWA for a specific point in time, namely, December 31, 2006. According to the labels in the process-produced data, we recoded the variables concurrently. The process-produced data differentiate between having or not having passed the Abitur (i.e., the upper secondary school completion certificate attesting aptitude for higher education), having completed or not having completed vocational training and the completion of a university of applied sciences or another college or university. The recoding resulted in 6 categories (see Table 1).

Table 1
Compared information on educational attainment

Comparable variable	Process-produced data
NOVOC =no Abitur without completed vocational training	0=Without completed education and vocational training (ohne Ausbildung) 1=Secondary/intermediate school completion certificate without completed vocational training (Volks-/Hauptschule, Mittlere Reife oder gleichwertige Schulbildung ohne Berufsausbildung) 21=No vocational training (ohne abgeschlossene Berufsausbildung)
VOC =no Abitur with completed vocational training	2=Secondary/intermediate school completion certificate with completed vocational training (Volks-/Hauptschule, Mittlere Reife oder gleichwertige Schulbildung mit Berufsausbildung) 22=Within-company vocational training/apprenticeship/traineeship (Betriebliche Ausbildung) 23=Extra-company (on-school) vocational training (Außerbetriebliche Ausbildung) 24=Specialized vocational school (full-time vocational school) (Berufsfachschule) 25=Technical school (Fachschule)
ABI =Abitur without completed vocational training	3=Upper secondary school completion certificate (general or subject-specific aptitude for higher education) without completed vocational training (Abitur/Hochschulreife (allgemein und fachgebunden) ohne Berufsausbildung)
ABIVOC =Abitur with completed vocational training	4=Upper secondary school completion certificate (general or subject-specific aptitude for higher education) with completed vocational training (Abitur/Hochschulreife (allgemein und fachgebunden) mit Berufsausbildung)
UAS =Completion of a university of applied sciences	5=Completion of a university of applied sciences (Abschluss einer Fachhochschule) 26=University of applied sciences (Fachhochschule)
UNI =College/university degree	6=College/university degree (Hochschul-/Universitäts-abschluss) 27=University (Hochschule/Universität)

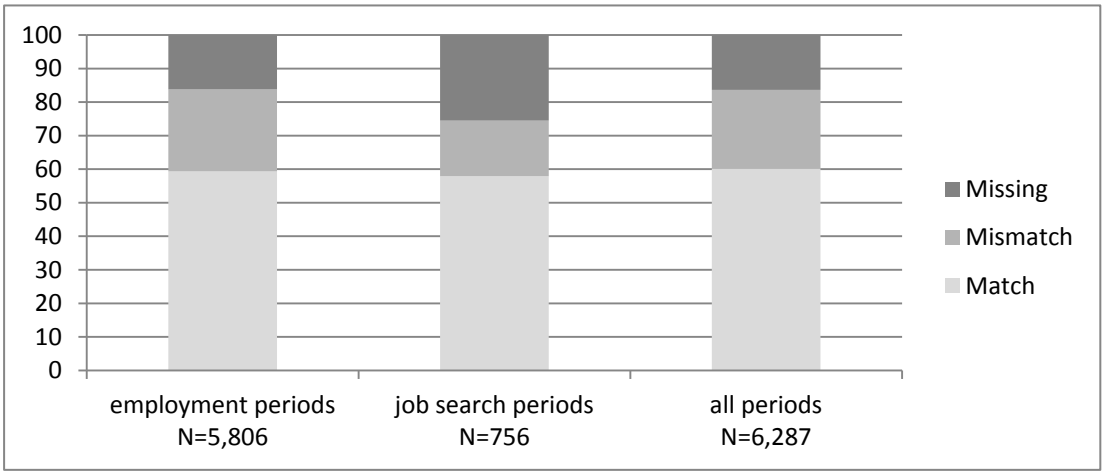
4 Structure of false and missing notifications

Let us turn to the questions of which groups are especially affected by missing or false notifications and which factors determine the probability of such failure. In the following, we apply descriptive statistics and conduct multinomial logistic regressions to answer these questions and to understand the structure of the false and missing values in the process-produced data.

The data used are based on the linked sample of the ALWA-ADIAB (8,166 respondents). To ensure high reliability of the educational information in the ALWA, foreign language interviews were dropped because those respondents were only asked about their educational degree instead of their educational history. To minimize errors concerning the reliability of the educational attainment information, we also excluded respondents whose life courses were not completely reported. Additionally, any ALWA respondent was dropped who did not have any episodes in the process-produced dataset at the selected time of comparison (December 31, 2006). As a result, there are 6,287 respondents in the dataset with information on educational attainment from the ALWA.

To describe the degree of the reliability problem in the dataset, we first descriptively looked at three constellations: A “match” means that the process-produced data reports exactly the same degree as the ALWA survey. A “mismatch” means that the process-produced data reports a different degree than the ALWA survey. “Missing” means that there is no information on educational attainment in the process-produced data to compare with the ALWA survey.

Figure 4
Reliability of educational attainment in SIAB data



Note: The sum of the total numbers of employment and job search periods reported here (N=6,562) does not correspond with the total number of persons (N=6,287) because there are respondents in the dataset who are employed while they are also searching for a job.

Source: ALWA-ADIAB, own calculations.

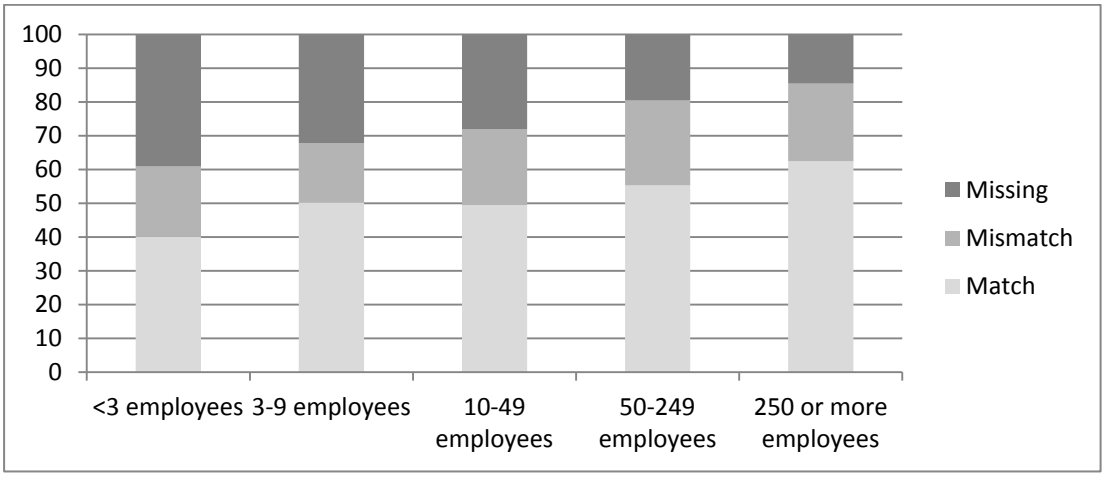
To check our first assumption, namely, that the educational information in employment episodes is less reliable than in job search episodes, we looked descriptively at the amount of matches, mismatches and missing data by comparing educational attainment in the ALWA with SIAB employment and SIAB job search periods (see Figure 4).

Looking first at the SIAB employment periods, it can be seen that nearly 60 percent of the respondents’ educational degrees match with reports in ALWA, almost 25 percent mismatch (meaning educational attainment is under- or overestimated), and more than 15 percent of the information is missing (see Figure 4, first bar).

Compared to using the educational information from the SIAB employment periods, one can say that the information from the SIAB job search episodes (see Figure 4, second bar) does not contain more reliable information about educational achievement: The number of matches is approximately the same as that identified by using the employment episodes, whereas the number of mismatches is lower and the amount of missing data is higher. Additionally, the low number of job search episodes also shows that only a few respondents are job seekers. Therefore, not enough information can be derived if only job search notifications are used. Looking at episodes of both types (see Figure 4, third bar), the picture – compared to that observed in the first bar – changes only a little bit. The reliability of the uncorrected SIAB data is at approximately 60 percent; about one-fourth of the respondents report a different level of educational attainment in the ALWA, and in more than 15 percent of the cases, we have no information about education in the SIAB data.

Looking at the same indicators using the SIAB with both data sources and differentiating according to firm size, we find that the reliability of information on educational achievement increases with firm size (see Figure 5).

Figure 5
Matches, mismatches and missing data by firm size



Source: ALWA-ADIAB, own calculations. N=6,287

The proportion of missing data is especially the highest in very small firms and diminishes continuously by firm size. We assume that in middle-sized or smaller firms, mandated tax consultants or similar service providers were often in charge of notifying the social security system. Because external tax advisors have only little relation to employees of firms for which they have to do the mandatory reporting, the reliability of educational achievement data is worse than at large firms. The greater reliability of educational information from large firms could be due to the availability of a personnel department that uses personnel files to give the correct notification information, which is often software-based.

As these descriptive observations show, data quality needs to be improved. Even the data from very large firms have a rather high percentage of missing and

mismatched data points. Data from job search periods are more reliable but are still not perfect and, more importantly, are not available for most people at a given point in time. For these reasons, there is a need to use correction rules to improve data quality.

5 Correction rules

Existing correction rules can be divided into different types: those that use the panel structure of within-person employment data, those that additionally use assumptions about the firm's reliability of notifications, and those that take into account the panel structure of the within-person information in both employment and job search data. In the following, we will describe these correction rules.

5.1 Using the panel structure of within-person employment data

The best-known and most often replicated form of improving the quality of process-produced information on education is to use the panel structure of the within-person employment data (Fitzenberger et al. 2006), hereafter referred to as Fitzenberger and colleagues named it: IP - Imputation Procedure.[3] There exist several versions of IPs, all of which make use of multiple notifications about the education of each employee (see section 2). In this way, they try to find out which information is most likely incorrect and to replace it with correct information that was reported before and/or after the wrong/missing information. The three basic logical assumptions are as follows: (1) a person's education can increase but not decline over the life course, (2) after entering work life, the educational degree usually stays the same and (3) employers have to report the employee's highest degree, not the one needed for the job.

The first version of their correction rules (IP1) only makes use of within-person employment information. The IP1-correction process of the education variable is structured into four steps: First, the extrapolation is prepared by setting all of the educational information for people under the age of 18 to "No Degree" and using only information from employment periods. In the second step, the actual extrapolation starts: degrees are (forward-) extrapolated to later periods in which no information or a lower degree is reported. This is done until the last period of employment is reached or until a period of higher education is reached. If either "Abitur without completed vocational training" (ABI) or "no Abitur with completed vocational training" (VOC) was reported in one period and the other was reported earlier, the new education information is set to "Abitur with completed vocational training" (ABIVOC) because those two cannot be ranked. In the third step, "backward extrapolation" is used to fill in remaining missing data points, which may

[3] Even though we know that imputation is not the precise name of these procedures and it would be more correct to speak of correction rules, we use the term that was used by the original authors.

occur at the beginning of an employment career before the first non-missing educational information. Because education is assumed to be constant over time, in this step, only information from later periods is extrapolated to earlier ones, taking into account certain age limits. In the last step, certain additional adjustments are made. If someone holds a qualified job ("Meister", "Facharbeiter", "Polier"), VOC is imputed if the reported degree is missing or lower than expected because this educational degree is usually (90%) reported for people holding those types of job. If there are parallel episodes in which someone holds more than one job at the same time, the highest reported educational achievement is extrapolated to all parallel periods (Fitzenberger et al. 2006). IP1 can be seen as the upper bound for educational achievement because there is no replacement of higher education with lower education, but there might be (false) replacement of lower education with higher education.[4]

Wichert and Wilke (2010) also made use of these correction rules, under the assumption that they deliver the best corrected data. Following IP1, they invented a similar correction procedure for the citizenship variable. They argue that both measures give reliable results if the errors are at random. After looking at the data from job search periods, they assume that the errors in the education variable are at random and that educational attainment notified in job search episodes can be used to test the effectiveness of the correction rules for employment periods. They compare job search periods with employment periods and find that job search periods are more reliable, although they are still erroneous. As they admit themselves, their "results may not hold for the entire German population" (Wichert and Wilke 2010: 11) because they only look at those employees who were recently looking for a job. We agree that this is a highly selective characteristic, and from our point of view, it definitely cannot be assumed that the subsample is representative of the entire German population. Although the authors are aware that this editing method only performs well in cases of error distribution at random, they still think it is appropriate because they "assume that the errors in the education and the nation variable can be considered as being random and therefore not deterministic error" (Wichert and Wilke 2010: 7). However, we do not agree with that statement, as our results will prove that there are various characteristics that influence the probability of false notification.

[4] Another version (IP2) of imputation by Fitzenberger et al. (2006) places some restrictions on which information can be extrapolated. IP2 is subdivided into versions A and B. Both versions try to minimize the possibility of false extrapolation. In IP2A, reliability is assumed only for any educational degree that is reported at least three times, except when education has only been reported four or fewer times in total. IP2B uses that rule only when the person's education sequence is inconsistent. Inconsistency is assumed when for a person a certain (lower) degree is reported after a higher one. Otherwise, IP2B extrapolates just as IP1. Missing information will be filled using extrapolation from valid spells.

5.2 Using the panel structure of within-person employment data and assumptions about the firm's reliability of notification

In another version, IP3, Fitzenberger and colleagues use the same procedure as described before but first divide employers into reliable and unreliable employers (Fitzenberger et al. 2006). In this version, it is not important how often a firm reports a certain degree for an employee. Only the structure of (changes in) reporting for one person counts. This is because unchanging reports may well be the results of copying the last reported degree without rechecking. Employers are considered reliable if they always report the same degree or if they change the reported degree only once. This is because the authors assume that it is highly unlikely that an employee would earn two new degrees while working at one workplace.

However, there are two types of changes in reports that IP3 would accept without downgrading a firm to 'unreliable'. One is the reporting of a lower degree than reported earlier if the lower degree is reported consistently from that point onwards. In this case, it is assumed that the employer wanted to report the lower degree from the beginning. The second exception is a change in the notifications that changed back immediately. If an employer reports a different degree only in one year and from the next year onwards reports the original degree, the employer does not lose its credibility.

The information on a firm's reliability is used in the following way: reports by unreliable employers are set as missing data, and information is only extrapolated from reliable employers (Fitzenberger et al. 2006: 419). The other extrapolation rules of IP3 are the same as IP1, which were described in the previous section (chapter 5.1.).

Drews (2006) tried to improve IP3 by using additional data to deduce a firm's reliability from its history of false notifications on the aggregate level. The basic idea was to categorize firms as reliable or non-reliable and to replace the unreliable firms' reports for all of its employees in the same manner as described above. This method was meant to increase the number of detected misreports and to thus create a more reliable dataset. However, it turned out that the results did not vary significantly from IP3.[5]

[5] Unfortunately, however, Drews did not describe his approach in much detail. Because there are various possibilities for deciding which firm should be treated as unreliable, it was not possible to replicate Drews' approach.

5.3 Using the panel structure of within-person employment as well as job search data

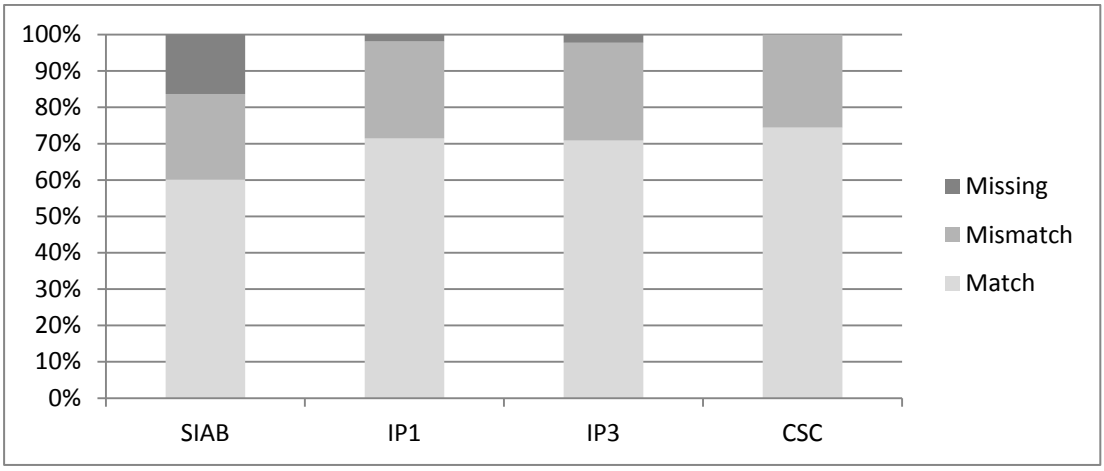
Taking into account employment and job search data, Kruppe (2006) developed a somewhat divergent way of correcting data to improve the quality of the education variable. Instead of only using information from employment episodes, he additionally takes information from job search episodes as a source. Making use of employment and job search data confers the advantage of more information. Because of the use of a combination of data sources, we call this version of correction “Combined Source Correction”, hereafter referred to as CSC.

In the first step of CSC, two variables are generated: one for school degree and one for vocational/university degree. For both variables, the available information from employment as well as job search episodes is used. If there is a discrepancy, it is assumed that the highest degree is valid. In the second step, missing data points were filled in by using educational information from the preceding period. The precondition for filling in those missing data points is that the periods are ordered chronologically. In the third step, for every point in time, the highest school degree and highest vocational/university degree are calculated by replacing any degree with the former one if that one is higher. This procedure is used for both variables separately. In the fourth step, a new variable is generated that contains the combined information on schooling as well as on the vocational/university degree. An additional fifth step has shown to be very effective: replacing the remaining missing data points with “no Abitur without completed vocational training” (NOVOC). The argument for this step is based on the assumption that, to reduce workload, the one who completes the form will (often) not fill in anything about a client without any educational degree or if low-qualified work is performed.

5.4 Effectiveness of correction rules: descriptive evidence

According to Fitzenberger et al. (2006: 408), a limitation of their “approach is that it is not possible to tell which one of the different imputed education variables is the best.” This is where we step in: How close is the information on educational attainment to that from the highly reliable ALWA data before and after the various corrections? First, let us have a descriptive look at the question of how suitable the correction rules are for improving data quality (see Figure 6).

Figure 6
Matches, mismatches and missing data by correction rule



Source: ALWA-ADIAB, own calculations. N=6,287

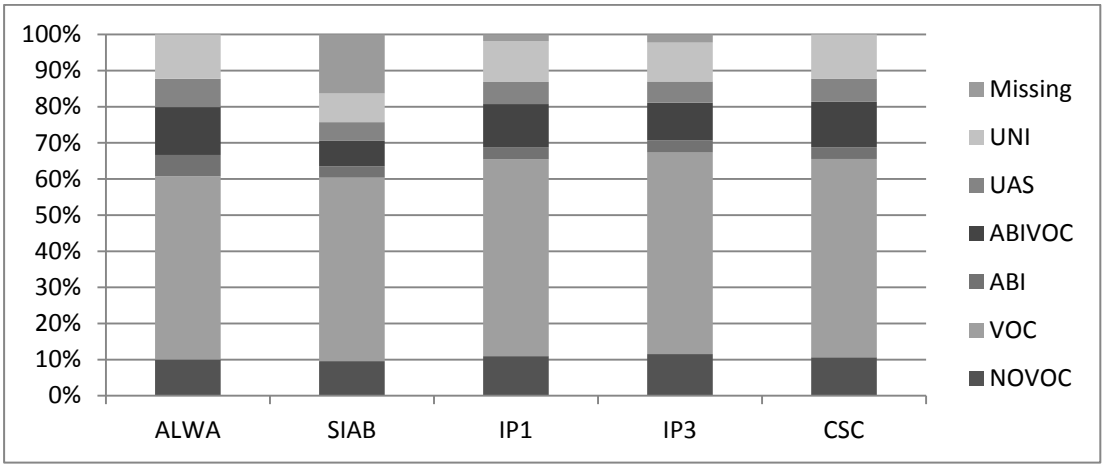
The first bar shows the correspondence between the information on educational achievement in the uncorrected data of the SIAB employment and job search dataset and the information from the ALWA. In the SIAB, only 60 % of educational information is correct. This number is significantly higher for each correction method. IP1 is a little more precise than IP3, with approximately 70 % matches. CSC turns out to identify a couple more educational degrees, ending up with approximately 75 % matches. The number of mismatches in all correction methods is nearly the same, but the number of missing data points can be reduced to a very small number by all correction methods and is eliminated with CSC. Although many missing data points get replaced with the correct information using CSC, a few are also replaced with false educational information.

This result leads to the questions of whether there is a structural bias in correcting educational information, and whether correction rules reduce or increase existing structural bias. We assume that reporting of the correct educational level of achievement of an employee depends on the characteristics of the person, the job and the employing firm. Therefore, we first analyze whether the educational structure is different after using the correction rules (see Figure 7).

Looking at the educational level of attainment “no Abitur without completed vocational training” (NOVOC), and comparing ALWA to SIAB as well as to the results using the correction rules, the SIAB underestimates and the correction rules overestimate the number of persons in this category slightly. In contrast, the SIAB is very good at reproducing the number of persons in the category “no Abitur with completed vocational training” (VOC), whereas all of the correction rules overestimate the number of persons with this level of educational attainment. Determining the correct number of persons holding an “Abitur without completed vocational training” (ABI) seems to be a very difficult task. SIAB as well as all of the correction rules underestimate the number of persons who fall into this special category. Perhaps this is the case because of difficulties in gathering accurate

information about the educational attainment of persons who achieve their secondary school completion degree after the regular school years. The category “Abitur with completed vocational training” (ABIVOC) is also confronted with this problem, but here, all of the correction rules are able to improve the data’s quality, with the highest improvement by CSC. The categories “Completion of a university of applied sciences” (UAS) and “College/university degree” (UNI) are corrected in the best possible way by CSC. Therefore, CSC seems to perform the best in minimizing the existing bias toward overestimating lower educational degrees in the SIAB.

Figure 7
Educational structure using different correction rules



Source: ALWA-ADIAB, own calculations. N=6,287

If we now look at the differences by gender, we can go into more detail about whether the correction rules can reduce the number of mismatches and missing data points for each educational level to the same extent. Negative numbers indicate an overestimation of the number of persons having a certain degree, and positive numbers point to an underestimation. In the table, bold numbers indicate the best performing correction measure for each degree (see Table 2).

Table 2 shows that the extent of underestimation without any correction (SIAB) is smallest for those males who do not hold an Abitur (NOVOC and VOC) and for those who passed the Abitur but did not complete any vocational training (ABI). The extent of underestimation is largest for higher degrees and missing data points. Except for ABI, this is also true for females. In general, the bias in the uncorrected data increases as the educational degree level increases.

Table 2
Differences in educational structure by gender

	Males (N=3,174)				Females (N=3,113)			
ALWA compared with								
	IEB	**IP1**	**IP3**	**CSC**	**IEB**	**IP1**	**IP3**	**CSC**
NOVOC	**-0.5%**	1.3%	2.0%	0.9%	**-0.4%**	0.5%	1.0%	**0.4%**
VOC	**1.8%**	3.2%	4.2%	3.8%	**-1.7%**	4.4%	6.0%	4.5%
ABI	**-2.3%**	-2.5%	-2.4%	-2.4%	-3.1%	-2.7%	**-2.6%**	-2.8%
ABIVOC	-5.5%	-1.3%	-2.5%	**-1.0%**	-7.0%	-1.3%	-3.1%	**-0.4%**
UAS	-2.5%	-2.0%	-2.1%	**-1.7%**	-2.8%	-1.5%	-2.0%	**-1.1%**
UNI	-4.3%	-0.3%	-1.0%	**0.5%**	-4.4%	-1.7%	-2.0%	**-0.6%**
missing	13.2%	1.4%	1.8%	**0.0%**	19.5%	2.4%	2.7%	**0.0%**

Source: ALWA-ADIAB, own calculations. N=6,287

Additionally, Table 2 shows that all correction rules begin to overcome the general underestimation of higher degrees. However, the best correction rule is CSC, particularly because it is the most precise in correcting data on higher degrees – regardless of the person's gender. Obviously, using educational information from both sources - employment as well as job search episodes - and simply replacing the remaining missing data points with NOVOC seems to be a very efficient method for correcting the information on educational achievement.

However, there is still a remaining difference, part of which is of substantial importance. Looking at the number of university degrees (UAS and UNI), CSC now slightly overestimates the educational attainment of males, whereas the percentage of females with a university degree remains underestimated, although they both start at approximately 4 % in the uncorrected version (SIAB).[6] This finding indicates that the errors in the educational achievement variable are not random. Therefore, in the following, we analyze the probabilities of matching the educational attainment from the SIAB and CSC to that reported in the ALWA, using multivariate logistic regression. The probabilities resulting from the uncorrected data on the one hand and from the same data after CSC correction on the other hand are compared within the same statistical model. CSC was chosen because it delivered the most effective version of a correction rule.

5.5 Effectiveness: Multivariate Evidence

The effectiveness of correction rules lies not only in their ability to match most often information on educational attainment as reported in the ALWA but also in their success in eliminating the structural bias that has been detected in the SIAB. Therefore, we ask whether CSC, as the best-performing correction rule, is able to overcome this structural bias using multivariate logistic regression. Table 3 shows the results based on 6,287 cases using uncorrected SIAB data (columns 2-4) and

[6] CSC indeed overestimates most often but underestimates least often at the same time and is particularly capable of overcoming the underestimation of the number of academically qualified women.

using CSC corrected data (columns 5-7). The variables used in the model are listed in column 1. We report odds ratios; these are the probabilities of matching the level of educational attainment reported in the ALWA. Missing data points and mismatches are the two non-match possibilities. A value of 1 or higher indicates that the probability of matching is higher than for the reference group, and a value lower than 1 indicates that the probability of matching is lower than for the reference group. Additionally, robust standard errors and significance levels are shown.

Columns 2 and 5 in Table 3 show that the probability of women's levels of educational attainment being reported correctly is lower than that of men's (males are the reference group). The effect of gender is significant for both models (columns 4 and 7). Even when including part-time employment as another variable, the gender effect remains significant. Surprisingly, part-time employment is not significant, but the interaction term of male gender and part-time employment is significant, which points to the particular proneness to error of atypical employment relationships. This (negative) effect of part-time work for males on the probability of correct reporting is only significant for the uncorrected data. This result means that CSC is able to overcome the underlying bias in gender-specific work-time effects.

Table 3
Probability of matching level of educational attainment as reported in the ALWA (Logistic Regression)

	SIAB			CSC		
	Odds Ratio	**Std. Err.**	**P>z**	**Odds Ratio**	**Std. Err.**	**P>z**
Gender Ref.: Male						
Female	0.81	0.06	0.00	0.82	0.06	0.01
Work time Ref.: Fulltime						
Part-time	0.97	0.08	0.69	0.95	0.08	0.52
Part-time*Male	0.71	0.10	0.02	0.79	0.12	0.12
Earnings per day Ref.: more than 110 €						
less than 20 €	0.72	0.08	0.00	1.32	0.15	0.02
20 to less than 60 €	1.24	0.12	0.03	1.32	0.14	0.01
60 to less than 110 €	1.34	0.11	0.00	1.25	0.12	0.02
Age Ref.: older than 45						
younger than 25	1.65	0.21	0.00	1.42	0.20	0.01
25 to younger than 35	1.25	0.12	0.02	1.29	0.14	0.01
35 to younger than 45	1.10	0.08	0.16	1.11	0.08	0.17
Duration of lifetime employment Ref.: more than 15 years						
less than 5 years	0.48	0.06	0.00	0.48	0.07	0.00
5 to less than 10 years	0.52	0.05	0.00	0.57	0.06	0.00
10 to less than 15 years	0.79	0.07	0.01	0.90	0.08	0.25
Firm size Ref.: more than 250 employees						
less than 3 employees	0.69	0.10	0.01	0.99	0.15	0.96
3 to less than 10 employees	0.84	0.08	0.07	1.14	0.12	0.22
10 to less than 50 employees	0.74	0.06	0.00	1.05	0.09	0.58
50 to less than 250 employees	0.79	0.06	0.00	1.01	0.08	0.93

(Continued)	SIAB			CSC		
	Odds Ratio	**Std. Err.**	**P>z**	**Odds Ratio**	**Std. Err.**	**P>z**
Nationality Ref.: Non-German						
German	1.29	0.19	0.09	1.68	0.26	0.00
Branch Ref.: Public administration						
Agriculture, forestry and fishing	0.77	0.22	0.35	0.64	0.19	0.13
Mining and quarrying	0.60	0.28	0.27	0.41	0.19	0.05
Manufacture of food products	1.21	0.27	0.40	1.81	0.50	0.03
Manufacture of apparel, jewelry and other articles	0.89	0.31	0.74	0.74	0.26	0.39
Manufacture of wood or paper and products of wood or paper	0.61	0.15	0.05	0.76	0.21	0.32
Publishing and printing	0.58	0.13	0.02	0.71	0.18	0.19
Manufacture of chemical products and other basic materials	1.10	0.19	0,60	1,04	0,20	0,82
Manufacture of fabricated metal products	0.75	0.15	0,15	1,05	0,23	0,81
Manufacture of machinery	0.84	0.13	0,27	0,84	0,14	0,29
Manufacture of motor vehicles	0.95	0.19	0,80	1,45	0,34	0,11
Water, electricity, gas and other suppliers and disposers	1.14	0.32	0,63	1,87	0,65	0,07
Construction and installation	0.93	0.17	0,70	0,91	0,18	0,65
Wholesale, retail trade and repair of motor vehicles	1.15	0.26	0,53	1,18	0,30	0,52
Wholesale	0.60	0.10	0,00	0,75	0,13	0,10
Retail trade	0.60	0.09	0,00	0,78	0,13	0,13
Hospitality and food service	0.61	0.12	0,01	0,66	0,14	0,05
Transporting and logistics	0.60	0.10	0,00	0,94	0,18	0,74
Financial, insurance activities	0.62	0.10	0,00	0,64	0,11	0,01
Renting and advertising	0.48	0.09	0,00	0,66	0,13	0,04
IT Services	0.45	0.10	0,00	0,66	0,16	0,08
Legal, tax and management consultancy	0.46	0.08	0,00	0,56	0,11	0,00
Engineering and related technical consultancy	0.68	0.14	0,05	0,84	0,18	0,41
Employment placement and agency	0.64	0.16	0,07	0,70	0,18	0,18
Security and cleaning service	0.58	0.12	0,01	0,87	0,20	0,56
Education	0.95	0.17	0,78	0,97	0,18	0,89
Health care	0.92	0.14	0,60	1,26	0,21	0,18
Social work	0.84	0.14	0,30	0,77	0,14	0,15
Societies and associations	0.64	0.14	0,04	0,77	0,18	0,28
Arts, recreation and sports	0.42	0.10	0,00	0,49	0,11	0,00
Private households	0.39	0.18	0,04	0,91	0,47	0,86
Constant	2.23	0.46	0,00	1,94	0,41	0,00
Number of obs.	6,287			6,287		
Wald chi2(47	414.6			175.5		
Prob. > chi2	0.000			0.000		
Pseudo R2	0.051			0.024		

Source: ALWA-ADIAB, own calculations

Compared to the reference category ‘daily earnings more than 110 Euro’, it is more likely that education will be reported correctly for those earning from ‘20 up to 110 Euro a day’ and less likely for earnings below 20 Euro. This finding indicates that firms do not bother with reporting the educational levels of low income workers correctly. Perhaps the effort of finding out and reporting the correct education level is quite large in proportion to the small cost factor that those workers’ wages constitute. However, it could also be the case that firms actually report the necessary educational degree for performing a job rather than the one that the employee actually holds. We suppose this is also a reason for the lower probability of accurate reports of educational attainment for persons earning more than 110 Euro a day. The effect is still significant in the second model, which means that CSC fails to correct this structural error.

Looking at age, one can say that the probability of accurate reporting decreases with age. Using the group ‘older than 45’ as a reference, there are significant effects for the age groups ‘younger than 25’ and ‘25 to younger than 35’. One possible explanation is that educational degrees are more relevant for placing younger employees at a firm, whereas for older employees, work experience is far more relevant. Another reason for the significant age effect could be that degrees that are completed after the first placement at a firm fail to be added to an employee’s personal file. Comparing both models shows that CSC additionally fails with respect to overcoming selectivity by correctly replacing respondents’ levels of educational attainment.

The probability of accurate reporting of education increases with the duration of lifetime employment. In reference to the respondents who have been employed for more than 15 years in total, the reports of all other groups are incorrect significantly more often. At first glance, this effect seems to contradict the previously reported age effect that the educational attainment of younger respondents is more often reported correctly than that of older ones. However, people with interrupted work histories in particular have been employed for a shorter time in total and might not find work that matches their actual qualifications to a greater extent as a result of their interrupted work histories. If employers tend to report the degree necessary for a certain job rather than the degree an employee actually holds, the educational attainment of overqualified employees might be underreported. Alternatively, such employees may have spent a good deal of time in school and may thus have many changes in their educational degrees that are not reported correctly by their firms. The significance of most of these variables in the CSC model indicates that even after using CSC correction rules, the selectivity of corrected notifications persists.

The significant effects of firm size show that employees in smaller firms have a lower probability of their level of educational achievement being correctly reported than those in companies of more than 250 employees. Because the effect of the firm

size variable becomes insignificant in the CSC model, we conclude that this simple editing rule is sufficient for eliminating the firm size effect.

German citizens have a significantly higher probability of their level of educational achievement being correctly reported. A reason for this might be the (false) assumption that people with migration background in general only have a lower education. Another reason is that ALWA reports foreign degrees as they are described by the surveyed people. Because for example some foreign degrees are not acknowledged as having the same status as the same degree achieved in Germany, they might be recognized as lower degrees, and therefore, those migrants might work in jobs that they are – by degree – overqualified for.

Division of employees according to the branch of the firm they work in shows that the bias in reporting educational attainment could be reduced for some branches but was increased for others (in reference to Public Service, which is assumed to be the most accurate one because of its highly developed reporting structure). The significant values less than 1 in some branches seems to be eradicated by CSC, possibly because many workers with low qualifications are employed in these branches and their educational degrees are corrected by CSC. That, by using CSC, some of the formerly insignificant branches become significant indicates that the correction procedure raises selectivity bias at the same time. A reason for these selective data improvements might again be that CSC best improves the reports of branches with a high number of workers with low qualifications, especially if there are many missing values. In addition, significant differences with respect to Public Service persist even after using CSC, which indicates that there are unobserved mechanisms behind the reporting of educational attainment that are uncorrectable by simple correction rules.

However, it is also important to note that the corrected R^2 is quite low. This observation further indicates that we do not know much about the processes behind accurate reporting. Perhaps a variable such as the position in which an employee works could be used to improve the quality of data when assuming that misreporting partially stems from employers who report the necessary level rather than the actual level of educational attainment. Additionally, another variable that delivers satisfying information about the qualifications necessary for a job or on employees' earnings per hour could help further in explaining the variance in the quality of educational reports. Nevertheless, it seems that there will never be enough information to perfectly correct the educational attainment variable only using simple correction rules. Therefore, further research should test whether multivariate imputation methods could improve data quality by also eliminating the non-random structural bias observed (Clogg et al. 1991).

6 Conclusion

The quality of process-produced data is usually considered to be correct. However, as has been noted by several authors, this is not always the case – especially not for non-target variables such as educational attainment. These authors have suggested various correction rules to improve data quality. Until now, it has not been possible to empirically test whether those deductive correction rules lead to an improvement in data reliability. This situation changed with the new dataset ALWA-ADIAB, which links survey data with process-produced data for the same individuals using personal identifiers. In this paper, we asked whether errors in administrative data could be ruled out using those simple correction rules.

Assuming that the ALWA survey data are of high reliability, we first looked descriptively at the matches, mismatches and missing data points between survey and process-produced data in the uncorrected data and again after using the correction rules. Specifically, we looked at whether a simple editing rule based on the idea that an individual's educational level cannot decrease over the life course improves the quality of the data substantially. CSC uses educational information from employment as well as job search episodes and simply replaces the remaining missing data points with an indicator of no vocational qualification. This approach turned out to be a very efficient method for correcting the information on educational achievement, indicating that some of the data problems that come with process-produced data can be solved using rather simple editing rules. However, because the errors in the educational achievement variable are not random, this correction rule is not sufficient to eliminate structural biases completely.

Gender, age and other important variables continue to be significant factors influencing the probability of misreporting and missing data even after using the best tested correcting measure, CSC. The results of the multivariate logistic regressions showed that there are quite a few significant variables still influencing the probability of accurate reporting even after correction rules have been applied. This result shows that there is bias in the structure of misreporting that cannot be eradicated using simple correction rules. Although some types of structural bias can be adjusted, others persist. This result shows that the errors in process-produced educational reports are not random, and thus, there is a need for the development of new imputation methods that are capable of dealing with a non-random error structure. Additionally, the estimations in this article only explain a very small percentage of the variance in both original and corrected process-produced data, indicating that imputation methods have to be developed to improve the structure of process-produced data.

None of the tested correction measures were able to eliminate errors sufficiently. Therefore, it is necessary to develop a more complex imputation method. This is an important task for future research. Only with reliable process-produced data is it possible to do significant research. Education, for example, is used as a control

variable for all types of social and economic research. However, if those data stem from process-produced data, they are most likely biased, and even the best existing correction measure cannot compensate for that yet. However, as long as there is no imputation method, a correction method such as CSC should be used.

References

Antoni, Manfred, Katrin Drasch, Corinna Kleinert, Britta Matthes, Michael Ruland, and Annette Trahms. 2010. Working and learning in a changing world. Part I: Overview of the study (FDZ Methodenreport, 05/2010; Second, updated version: March 2011). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Antoni, Manfred and Stefan Seth. 2012. "ALWA-ADIAB - Linked individual survey and administrative data for substantive and methodological research." Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 132:141-146.

Baur, Nina. 2011. "Mixing process-generated data in market sociology." Quality & Quantity 45:1233-1251.

Benitez-Silva, Hugo, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust. 2004. "How large is the bias in self-reported disability?" Journal of Applied Econometrics 19:649-670.

Bernhard, Sarah, Christian Dressel, Bernd Firzenberger, Daniel Schnitzlein, and Gesine Stephan. 2006. "Überschneidungen in der IEBS: Deskriptive Auswertung und Interpretation." in FDZ Methodenreport.

Bollinger, Christopher R. and Martin H. David. 2005. "I didn't tell, and I won't tell: dynamic response error in the SIPP." Journal of Applied Econometrics 20:563-569.

Büttner, Thomas and Susanne Rässler. 2008. "Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity." IAB Discussion Paper 44.

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidmane. 1991. "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression." Journal of the American Statistical Association 86:68-78.

Davies, Paul S. and T. Lynn Fisher. 2008. "Measurement Issues Associated with Using Survey Data Matched with Administrative Data from the Social Security Administration." Proceedings of the American Statistical Association, Section on Government Statistics:147-157.

Dorner, Matthias, Jörg Heining, Peter Jacobebbinghaus, and Stefan Seth. 2010. "The Sample of Integrated Labour Market Biographies." Schmollers Jahrbuch 130:599-608.

Drasch, Katrin and Britta Matthes. 2011. "Improving retrospective life course data by combining modularized self-reports and event history calendars. Experiences from a large scale survey." Quality and Quantity. International Journal of Methodology Online First:1-22.

Drews, Nils. 2006. "Qualitätsverbesserung der Bildungsvariable in der IAB Beschäftigtenstichprobe 1975-2001." FDZ Methodenreport 05:1-16.

Fitzenberger, Bernd, Aderonke Osikominu, and Robert Völter. 2006. "Imputation Rules to Improve the Education Variable in the IAB Employment Subsample." Zeitschrift für Wirtschafts- und Sozialwissenschaften 126:405-436.

Gottschalk, Peter and Minh Huynh. 2010. "Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error." Review of Economics and Statistics 92:302-315.

Hethey-Maier, Tanja and Stefan Seth. 2010. Das Betriebs-Historik-Panel (BHP) 1975-2008. Handbuch Version 1.0.2. (FDZ Datenreport, 04/2010). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Hochfellner, Daniela, Dana Müller, and Anja Wurdack. 2012. "Biographical data of social insurance agencies in Germany. Improving the content of administrative data." Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 132:443-451.

Huber, Martina and Alexandra Schmucker. 2009. "Cleansing Procedures for Overlaps and Inconsistencies in Administrative Data. The Case of Length of Unemployment in German Labour Market Data." Social Bookkeeping Data: Data Quality and Data Management (Special Issue of Journal Historical Social Research) 34:230-241.

Jaenichen, Ursula, Thomas Kruppe, Gesine Stephan, Britta Ullrich, and Frank Wießner. 2005. "You can split it if you really want: Korrekturvorschläge für ausgewählte Inksonsistenzen in IEB und MTG." FDZ Datenreport.

Johansson, Per and Per Skedinger. 2009. "Misreporting in register data on disability status: evidence from the Swedish Public Employment Service." Empirical Economics 37:411-434.

Johnson, Barry W. and Kevin Moore. 2008. "Differences in Income Estimates Derived from Survey and Tax Data." Pp. 1495–1503 in Proceedings of the American Statistical Association, Section on Survey Research Methods.

Kapteyn, Arie and Jelmer Y. Ypma. 2007. "Measurement Error and misclassification. A comparison of survey and register data." Journal of Labor Economics 25:513-551.

Kleinert, Corinna, Britta Matthes, Manfred Antoni, Katrin Drasch, Michael Ruland, and Annette Trahms. 2011. "ALWA - New life course data for Germany." Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 131:625-634.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010. "Nonresponse and measurement error in employment research. Making use of administrative data." Public Opinion Quarterly 74:880-906.

Kröger, Katharina, Uwe Fachinger, and Ralf K. Himmelreicher. 2011. Empirische Forschungsvorhaben Zur Alterssicherung. Einige Kritische Anmerkungen Zur Aktuellen Datenlage (RatSWD Working Paper No. 170).

Kruppe, Thomas. 2006. Die Förderung beruflicher Weiterbildung - eine mikroökonometrische Evaluation der Ergänzung durch das ESF-BA-Programm (IAB Discussion Paper 21/2006). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Kruppe, Thomas. 2009. "Empirical Consequences of Defintions. The Case of Unemployment in German Register Data." Social Bookkeeping Data: Data Quality and Data Management (Special Issue of Journal Historical Social Research) 34:138-148.

Lyberg, Lars E., Paul Biemer, Martin Collins, Edith D. De Leeuw, Cathryn Dippo, Norbert Schwarz, Dennis Trewin, and (Eds.). 2012. Survey Measurement and Process Quality. New York u.a.: John Wiley & Sons.

Matthes, Britta and Annette Trahms. 2010. Working and Learning in a Changing World. Part II: Codebook (FDZ Datenreport, 02/2010). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Røed, Knut and Oddbjørn Raaum. 2003. "Administrative registers – Unexplored reservoirs of Scientific Knowledge?*." The Economic Journal 113:258-281.

Schnell, Rainer. 2011. *Survey-Interviews: Methoden Standardisierter Befragungen (Studienskripten zur Soziologie)*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Scioch, Patrycja. 2010. "The impact of cleansing procedures for overlaps on estimation results. Evidence for German administrative data." *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften* 130:485-512.

Scioch, Patrycja and Dirk Oberschachtsiek. 2009. "Cleansing Procedures for Overlaps and Inconsistencies in Administrative Data. The Case of German Labour Market Data." *Social Bookkeeping Data: Data Quality and Data Management (Special Issue of Journal Historical Social Research)* 34:242-259.

Seysen, Christian. 2009. "Effects of Changes in Data Collection Mode on Data Quality in Administrative Data. The Case of Participation in Programmes Offered by the German Employment Agency." *Social Bookkeeping Data: Data Quality and Data Management (Special Issue of Journal Historical Social Research)* 34:191-203.

Wagner, Joachim. 2012. *Daten des IAB-Betriebspanels und Firmenpaneldaten aus Erhebungen der Amtlichen Statistik – substitutive oder komplementäre Inputs für die Empirische Wirtschaftsforschung? (Working Paper Series in Economics No. 252)*. Lüneburg: University of Lüneburg.

Wichert, Laura and Ralf A. Wilke. 2010. *Which factors safeguard employment? An analysis with misclassified German register data. (FDZ Methodenreport, 11/2010)*. Nürnberg: FDZ.

Recently published

No.	Author(s)	Title	Date
21/2013	Singer, Ch. Toomet, O.-S.	On government-subsidized training programs for older workers	12/13
22/2013	Bauer, A. Kruppe, Th.	Policy Styles: Zur Genese des Politikstilkonzepts und dessen Einbindung in Evaluationsstudien	12/13
1/2014	Hawranek, F. Schanne, N.	Your very private job agency	1/14
2/2014	Kiesl, H., Drechsler, J.	Beat the heap	2/14
3/2014	Schäffler, J., Hecht, V. Moritz, M.,	Regional determinants of German FDI in the Czech Republic	2/14
4/2014	Prantl, S. Spitz-Oener, A.	Interacting product and labor market regulation and the impact of immigration on native wages	2/14
5/2014	Kohlbrecher, B. Merkl, C. Nordmeier, D.	Revisiting the matching function	2/14
6/2014	Kopf, E., Zabel, C.	Active labour market programmes for women with a partner	2/14
7/2014	Rebien, M., Kubis, A., Müller, A.	Success and failure in the operational recruitment process	3/14
8/2014	Mendolicchio, C. Pietra, T.	On the efficiency properties of the Roy's model under asymmetric information	3/14
9/2014	Christoph, B. Pauser, J. Wiemers, J.	Konsummuster und Konsumarmut von SGB-II-Leistungsempfängern	4/14
10/2014	Bossler, M.	Sorting within and across establishments	4/14
11/2014	Gillet, H. Pauser, J.	Efficiency in public input provision in two asymmetric jurisdictions with imperfect labour markets	4/14
12/2014	Antoni, M. Janser, M. Lehmer, F.	The hidden winners of renewable energy promotion	5/14
13/2014	Müller,S. Stegmaier, J.	Economic failure and the role of plant age and size	5/14
14/2014	Gärtner, D. Grimm, V. Lang, J. Stephan, G.	Kollektive Lohnverhandlungen und der Gender Wage Gap	5/14

As per: 2014-05-20

For a full list, consult the IAB website
http://www.iab.de/de/publikationen/discussionpaper.aspx

Imprint

IAB-Discussion Paper 15/2014

Editorial address
Institute for Employment Research
of the Federal Employment Agency
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Regina Stoll, Jutta Palm-Nowak

Technical completion
Jutta Palm-Nowak

Website
http://www.iab.de

Download of this Discussion Paper
http://doku.iab.de/discussionpapers/2014/dp1514.pdf

ISSN 2195-2663

For further inquiries contact the author:

Thomas Kruppe
Phone +49.911.179 5649
E-mail thomas.kruppe@iab.de

Britta Matthes
Phone +49.911.179 3074
E-mail britta.matthes@iab.de