Schmollers Jahrbuch 134 (2014), 237 – 248
Duncker & Humblot, Berlin

European Data Watch

This section offers descriptions as well as discussions of data sources that are of interest to social scientists engaged in empirical research or teaching courses that include empirical investigations performed by students. The purpose is to describe the information in the data source, to give examples of questions tackled with the data and to tell how to access the data for research and teaching. We focus on data from German speaking countries that allow international comparative research. While most of the data are at the micro level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) are included as well. Suggestions for data sources to be described in future columns (or comments on past columns) should be send to: Joachim Wagner, Leuphana University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@leuphana.de⟩. Past "European Data Watch" articles can be downloaded free of charge from the homepage of the German Council for Social and Economic Data (RatSWD) at: http://www.ratswd.de.

Linked Employer-Employee Data on Firms' Training Costs: Enriching Register based LEE Data with Firm Level Data on Apprenticeship Training

By Hans Dietrich, Holger Alda, Harald Pfeifer, Felix Wenzelmann, Gudrun Schönfeld, Stefan Schiel and Stefan Seth

1. Introduction

Despite the general trend towards tertiary education in Germany, more than 60% of each age cohort still enrols in apprenticeship training programs (Bundesministerium für Bildung und Forschung, 2009). Apprenticeship training, thus, remains the dominant educational pathway at the upper secondary level.

An important feature of apprenticeship training is the strong link between education and the labour market because a large part of the training takes place

within firms. A considerable share of former apprentices are retained by the training firms, and this facilitates the transition of trainees from education to the labour market (for statistics on retention rates see Hartung, 2013).

With respect to such labour market transition from an individual perspective, several data sets exist for research purposes. However, data sets representing the firm perspective are limited in their number, and they face severe limitations. The BIBB studies on costs and benefits of apprenticeship training (in the following BIBB CBS) are examples of surveys aimed at filling this gap (see Noll et al., 1983, von Bardeleben et al., 1995, Beicht et al., 2004; Schönfeld et al., 2010).

The latest survey of the BIBB CBS was conducted in 2008, reporting training costs and benefits for the reference year 2007. A total of 2,986 firms were interviewed that provided training in 51 occupations. Because the sample was drawn from the Establishment Register (*Betriebsdatei*) of the Federal Employment Agency (*Bundesagentur für Arbeit, BA*), the survey data can be merged with register data using an establishment-specific identifier, which, in most cases, corresponds with a firm as a local unit (more in detail see Alda et al. 2013). Technically, the data merge is feasible for all firms observed in the BIBB CBS 2007. However, a precondition for the merging procedure is that firms have formally agreed to the data merge during the survey interview. Two thirds of all of the firms included in CBS 2007 agreed with the data match.

The remainder of the paper is structured in the following way. Section 2 first describes the two data sources used for the merge and then addresses potential selectivity problems arising from a firm's option to decline the data merge. Section 3 discusses the quality of the match between survey and administrative data. Section 4 briefly summarizes the main results of the paper and sketches research perspectives of the matched data.

2. Data Sources of the Linked Employer-Employee Data

2.1 The Cost-Benefit Survey 2007

Based on a comprehensive report of the *Sachverständigenkommission Kosten und Finanzierung der Beruflichen Bildung* from 1974 (Sachverständigenkommission, 1974) the BIBB Cost and Benefit Surveys were conducted to obtain representative data for the costs and the benefits of apprenticeship training in (West-)Germany. Later on, the surveys were designed to deliver cost and benefit information in a structure which enables differentiation of costs and benefits along structural variables, such as economic sector, occupation, firm size or region. Finally, all cost benefit surveys focused on specific training occupations among all of the approximately 320 regulated training occupations in Germany. For the BIBB CBS survey of 2007, 51 of these occupations were

chosen based on their representativeness and practical relevance. The 51 occupations represent approximately 70% of all German apprenticeship participants.

The survey of the year 2007 includes detailed information about the firm's costs of training. These include monetary measures of expenditures for apprenticeship wages, costs for (full- and part-time) trainers and costs for training infrastructure (e.g., training facilities and material). For the calculation of the net costs, the studies also measure the benefits of training in the form of productive work performed by apprentices. Finally, the survey inquires about the firm's cost of recruiting already trained workers from the external labour market. Detailed results of the most recent BIBB CBS 2007 are published in Schönfeld et al. (2010). The BIBB CBS results have been extensively used for theoretical and empirical works in the field of education economics (see, e.g., Acemoglu/Pischke, 1998 or Mühlemann et al., 2010).

For the BIBB CBS 2007, a sample of firms was drawn from the BA-Establishment register. The register contains all firms with at least one employee subject to social security payments. Because all firms fitting this criterion are obliged to register, the representativeness of this pool of addresses can be taken as given. A gross sample of 22,000 firms was drawn, which have provided training in at least one of the 51 occupations of interest in June 2006, the reference period for sample selection.

Because the Establishment Register contains occupational information only on a 3-digit level, the survey institute had implemented a screening procedure to ensure that the respective firm was actually providing training in the chosen 4-digit occupation of interest. Approximately 60% of all 22,000 firms provided training in the chosen training occupation. The remaining 40% either did not train in the specific 4-digit occupation or, due to the time lag between sampling procedure and reference period of the interview, did not train apprentices at all. After the screening, approximately 13,000 addresses of the training firms of interest were identified. Field managers used 8,907 of these to conduct the intended number of 3,000 valid interviews. The survey was conducted between April and September 2008, and the average interview duration was 72 min.[1] The interview partner was the person responsible for the organisation of training, and/or the firm's human resource manager. In small firms, the interview was conducted with the owner or managing director of the firm. The field work was administered by infas (Institute for Applied Social Sciences), Bonn.

The survey institute infas established several techniques to raise data quality. For example, infas defined confidence intervals for several key variables (e.g., wages and training times) in the CAPI questionnaire. In case the interviewed person stated values outside the confidence interval, the CAPI program looped

[1] For more details see infas (2008) and Schönfeld et al. (2010).

back to the value and the interviewer had to ensure that the interview partner understood the question correctly. In case of incomlete answers for these crucial variables, an ex-post imputation technique was used to replace missing values. For most of the variables, missing values made up less than 5% of all answers, with somewhat higher shares for wage information on different groups of employees (10% to 15%). For detailed information about the imputation procedures see Alda (2010).

The merged data set, which is described in this paper, contains records from those firms that agreed to the merging of the firm specific survey data with the BA register data. Of the 2,986 firms, 2,083 (i.e., almost 70%) provided permission for the merge. Textbox A1 quotes the respective question from BIBB CBS 2007 and Table A1 provides an overview of the topics covered by the survey[2]. Table A2 displays descriptive information for the main variables for the subsample of merged firms and for the full sample of firms.

In order to test whether firms agreeing to the merge differ systematically from those refus the merge, the decision variable (1 for permission, 0 for refusal) was regressed on a set of structural variables, such as firm size, economic sector, region and training occupation (see Table 1). The results show that neither firm size indicators nor the variables of economic activity or region significantly alter or reduce the probability of belonging to the group of firms that did not permit the data matching. Furthermore, none of the dummy variables for the 51 observed occupations are significant at the 5% level. The observed low Pseudo R^2, of barely 0.02 shows that very little of the variance in the firm's merge permission is explained by the set of structural variables. The probit estimates thus lead to the conclusion that the two groups of firms do not differ systematically with respect to (observed) structural characteristics.

Table 1

Probit regression – dependent variable: firms' permission to merge data

	Model 1		Model 2		Model 3		Model 4	
	coef	se	coef	se	coef	se	coef	se
Firm-size:								
Ref. 50–499 employees								
1–9 employees	0.11*	0.063	0.110*	0.064	0.110*	0.064	0.060	0.079
10–49 employees	0.13**	0.062	0.137**	0.062	0.137**	0.062	0.105	0.068
500 and more employees	–0.11	0.096	–0.129	0.097	–0.130	0.097	–0.125	0.100

2 More detailed information about the set of firm level variables available is given in Alda (2010).

Economic sector: Ref. Public administration, education and health								
Manufacturing, agriculture, mining and quarrying, construction			–0.082	0.079	–0.082	0.079	0.007	0.120
Trade			–0.102	0.093	–0.102	0.093	–0.026	0.137
Services I			–0.037	0.087	–0.037	0.087	0.048	0.115
Services II			0.029	0.101	0.029	0.101	0.022	0.139
Region: Ref. East Germany								
West Germany					–0.005	0.054	0.015	0.055
Occupation:	No		No		No		Yes	
Pseudo R2	0.002		0.003		0.003		0.021	
Constant	0.45***	0.045	0.501***	0.077	0.508***	0.102	0.315	0.279

Source: BIBB CBS 2007. Level of significance: *** p<0,01, ** p<0,05, * p<0,1. N=2986.

2.2 Establishment History Panel (BHP) and the Integrated Employment Biographies (IEB)

The second part of the data is derived from the BA register data. Here, two types of process produced data are available from the 2007 survey records of firms that permitted data matching: a) firm data from the Establishment History Panel (BHP) and b) employee data from the Integrated Employment Biographies (IEB). The Establishment History Panel (BHP) is composed of cross sectional data sets since 1975 for West Germany and 1991 for East Germany. Every cross section contains all of the establishments in Germany that are covered by the Employment History (BeH) on June 30th. These are all establishments with at least one employee liable to social security on the reference date. Since 1999 establishments with no employee liable to social security but with at least one marginal part-time employee are also included. The cross sections can be combined to form a panel. The BHP contains information about the branch of industry and the location of the establishment. Furthermore, the number of employees liable to social security is also included. In addition, marginal part-time employees are also included (since 1999), both in total numbers and broken down by gender, age, occupational status, qualification and nationality. Quartiles of ages and wages are also provided, both for full-time employees alone as well as for all employees (see Spengler, 2008 and http://fdz.iab.de/en/FDZ_Establishment_Data/Establishment_History_Panel.aspx).

The Integrated Employment Biographies (IEB) consists of all individuals in Germany, who are characterized by at least one of the following employment statuses: employment subject to social security (in the data since 1975); marginal part-time employment (in the data since 1999); benefit receipt according

to the German Social Code III or II (SGB III since 1975, SGB II since 2005); officially registered as job-seeking at the German Federal Employment Agency; (planned) participation in programs of active labour market policies (in the data since 2000). The information delivered from different data sources are consolidated in the IEB. Each employment status is represented on a daily basis (see Dorner et al., 2010).

The register data enables the generation of variables that identify the employment background of the apprentice before the start of the apprenticeship, such as prior training qualifications, unemployment durations and schooling level (graduation from high school, *Abitur*). Register data also provide detailed information regarding the training period itself, such as the apprentice's training pay, the duration of training and whether the individual received social benefits during the training phase. Finally, the register data contain information about the apprentice's labour market integration after training.

3. Merging Quality

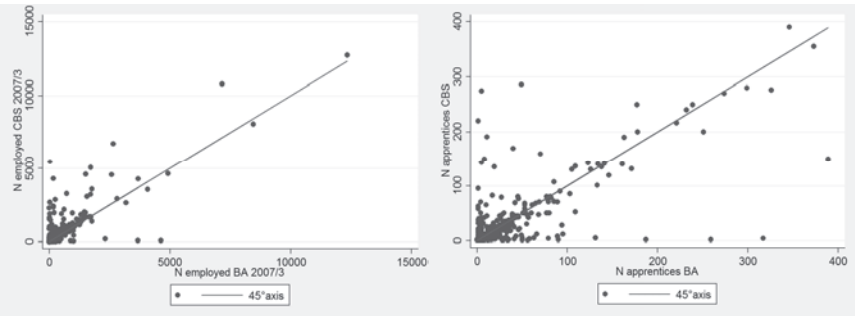
Although the merging of survey data and register data is technically trivial due to the unique firm identifier, the quality of the merge needs to be assessed on the basis of statistical analysis. Several mechanisms may lead to poor quality of the match even if technically the same firm is identified in both data sources:

First, different reference periods and/or a reference basis (e.g., mixing up marginal employment with part-time employment) could lead to diverging information about the firm size. Second, interview partners may have difficulties in remembering the correct numbers or values for the staff working in the firm. Third, the firm's administrative identification number may have changed over time. Although identification numbers are usually adjusted accordingly, a change in the identification number could occur in cases of in- or outsourcing or a change of ownership. Fourth, interview partners may only have a rough idea about which parts of the firm are covered by the administrative identification number, which can lead to under- or overestimation of the values delivered by the register data.

To assess the quality of the firm-level merge, we compare the firm's number of employees and apprentices reported in each of the two data sources. The number of employees covers all full- and part-time employees. Excluded from both data sources are employees with a marginal employment contract below the tax and social security thresholds (*geringfügig Beschäftigte*)[3]. The number

[3] Furthermore, those employees who are not subject to social security payments are not included in the employee variables. These are, for example, freelancers, public servants and short-term trainees.

of apprentices includes those apprentices for whom the interviewed firm contributes to social security. Apprentices who are (temporarily) trained by the reporting firm but are *not* subject to social security payments are not included.[4] Figure 1 presents scatter plots of the calculated differences. The plots indicate that the large majority of firms show no or small differences in their reported numbers of employees (left panel) and apprentices (right panel).



Source: BIBB CBS and BHP (30. September 2007).

Figure 1: Scatter plot of differences
in the number of employees and apprentices

To be able to differentiate between different merge quality levels, we introduce categories for the quality of fit. We differentiate between a) a perfect fit, b) a good fit (using firm size adjusted confidence bands) and c) a poor fit of the numbers of employees and apprentices reported both in survey and register data. A "perfect fit" is identified when the respective number of employees and apprentices does not diverge at all. A "good fit" is identified in the case of employees if the difference between the number of employees reported by the BIBB CBS and the number of employees reported in the BA data is smaller than half of the respective firm-size category width. In the case of apprentices, a "good fit" refers to a deviation between the two data sources of between 1 and 4 apprentices. A "poor fit" refers to firms outside the thresholds defined for the "good fit". Finally, we report d) the firms with missing data on employment for which the merge quality cannot be assessed (for a more detailed presentation of the quality assessment see Alda et al., 2014).

Table 2 reports the results of the quality assessment. Close to 90% of all merged firms fit perfectly or fit well with respect to their reported numbers of employees and apprentices. For approximately 10% of the merged firms, we find a poor merge quality, while 1% (in the case of employees) and 0.2% (in

4 It can be the case that the interviewed firm takes over (parts of) the training for other firms or other organizational units that are not covered by the firm identifier.

the case of apprentices) are not classified due to problems with missing data (Table 2).

Table 2

Quality of fit between survey- and register data based on the number of employees and apprentices

	Employees		Apprentices	
	Share of firms (%)	Number of firms	Share of firms (%)	Number of firms
a) Perfect fit	10.4	207	55.9	1117
b) Good fit	78.2	1561	33.2	664
c) Poor fit	10.0	200	10.7	213
d) Missing data	1.5	29	0.2	3
Total	**100**	**1997**	**100**	**1997**

Source: BIBB CBS and BHP (30. September 2007).

Although we find a good fit for both the number of employees and apprentices, a remarkably higher proportion of perfect fits (= identical figures) can be seen in the apprenticeship figures. One reason for this marked difference in the share of perfect fits can be due to the fact that some groups, such as marginal employees, temporary work agency employees, family helpers or proprietors could be reported in the survey but are not included in the number of employees as measured in the administrative data. Furthermore, the number of employees is much more volatile due to labour turnover, when compared to the number of apprentices in the firm. Finally, the person interviewed in the firm was requested to be informed about the costs and the benefits of apprenticeships. Due to this focus of the survey, we can expect more accurate reporting of apprenticeship numbers than of employment numbers. However, for both groups, we find that approximately 10% of the firms diverge strongly with respect to the numbers reported in the survey and in the register based data. The variable identifying the quality of the data fit will be included in the data file accessible at the Research Data Center of the BA in the IAB (FDZ).

4. Summary and Research Perspectives

The aim of the paper was to describe the merging of BIBB survey data and IAB process-generated (administrative) data. We discuss potential selectivity problems arising due to the option of firms to decline the data merge. Using a probit model including standard structural control variables, we find no indica-

tion of structural differences between the group of firms declining and the group of firms agreeing to the data match. The second part of the paper presents results of a data merge quality assessment. Comparing the number of firms' employees and apprentices reported in the two data sources, we find a sufficient merge quality for approximately 90% of the firms. For only approximately 10% of the firms, we find substantial differences in these reported numbers.

This exercise provides a brief example of how firm-level survey information can be used to increase the analytical potential of administrative data. Because the BIBB CBS 2007 contains detailed information about training costs, training benefits and training strategy of firms, the potential for researchers interested in the fields of vocational training and labour market integration is large. For example, the BIBB survey measures the average productivity of apprentices compared to already trained workers in the firm. Depending on the productivity level of apprentices, employment, retention, occupational mobility and also post-training wages might vary. Such and several other questions can be addressed with the new BIBB-IAB data set, which will be available in the Research Data Center of the German Federal Employment Agency (Bundesagentur für Arbeit) at the IAB.

References

Acemoglu, D./*Pischke*, J.-S. (1998): Why Do Firms Train? Theory and Evidence. Quarterly Journal of Economics 113 (1), 79–119.

Alda, H. (2010): Kosten und Nutzen der betrieblichen Berufsausbildung in Deutschland – Beschreibung der Datensätze für die Jahre 2000 und 2007, BIBB-FDZ Daten- und Methodenberichte 7.

Alda, H./*Dietrich*, H./*Pfeifer*, H./*Wenzelmann*, F. (2014): Verknüpfungsqualität der Surveydaten „Kosten und Nutzen der betrieblichen Berufsausbildung" aus dem Jahr 2007 mit Registerdaten zur Beschäftigung, BIBB-FDZ Daten- und Methodenberichte (forthcoming).

Bardeleben, R. v./*Beicht*, U./*Fehér*, K. (1995): Betriebliche Kosten und Nutzen der Ausbildung. Repräsentative Ergebnisse aus Industrie, Handel und Handwerk, Bielefeld.

Beicht, U./*Walden*, G./*Herget*, H. (2004): Kosten und Nutzen der betrieblichen Berufsausbildung in Deutschland, Bielefeld.

Bundesministerium für Bildung und Forschung (2009): Berufsbildungsbericht 2009, Bonn/Berlin.

Dorner, M./*Heining*, J./*Jacobebbinghaus*, P./*Seth*, S. (2010): The Sample of Integrated Labour Market Biographies. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 130 (4), 599–608.

Hartung, S. (2013): Ausbildungsberechtigung, Ausbildungsaktivität und Übernahmeverhalten von Betrieben, in: Bundesinstitut für Berufsbildung (eds.), Datenreport zum

Berufsbildungsbericht 2013. Informationen und Analysen zur Entwicklung der beruflichen Bildung, Bonn.

Infas (2008): Betriebsbefragung zu den Kosten und dem Nutzen der betrieblichen Berufsausbildung. Bonn (Infas).

Mühlemann, S. / *Pfeifer*, H. / *Walden*, G. / *Wenzelmann* F. / *Wolter*, S. C. (2010): The Financing of Apprenticeship Training in the Light of Labor Market Regulations. Labour Economics 17, pp. 799–809.

Noll, I. / *Beicht*, U. / *Boll*, G. / *Malcher*, W. / *Wiederhold-Fritz*, S. (1983): Nettokosten der betrieblichen Berufsausbildung, Berlin.

Sachverständigenkommission Kosten und Finanzierung der beruflichen Bildung (1974): Kosten und Finanzierung der außerschulischen beruflichen Bildung. Abschlussbericht, Bielefeld.

Schönfeld, G. / *Wenzelmann*, F. / *Dionisius*, R. / *Pfeifer*, H. / *Walden*, G. (2010): Kosten und Nutzen der dualen Ausbildung aus Sicht der Betriebe. Ergebnisse der vierten BIBB-Kosten-Nutzen-Erhebung, Bielefeld.

Spengler, A. (2008): The Establishment History Panel, Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 128 (3), 501 – 509.

Appendix

Table A1

Cost-Benefit Survey 2007 – Overview of main themes and variables

General information about the firm concerning legal status and the institutional framework in the firm, such as employee representation in the workplace and collective agreement coverage
Information about firms' current and expected employment and qualification structure
Information about the gross cost components of apprenticeship training in the chosen occupation, such as: • Costs for trainers • Costs for training administration • Costs for training infrastructure
Information about the productive work of apprentices (incl. equivalence benefits) on the level of unskilled and skilled workers
Information about internal training centres and in-house classroom training
Information about reimbursement/ financial support
Non-monetary benefits/ advantages of apprenticeship training
Information on market strategy, product regime and organizational changes
Subjective appraisals/ company measures relating apprenticeship training

Table A2

Descriptive sample information

Firm size	Sample match permission	Full sample of firms
1 – 9 employees	0.54	0.54
10 – 49 employees	0.33	0.33
50 – 499 employees	0.12	0.12
500 and more employees	0.01	0.01
Region		
West Germany	0.83	0.83
East Germany	0.17	0.17
Training sector		
Trade and industry	0.48	0.48
Crafts and skilled trades	0.29	0.31
Agriculture	0.03	0.03
Liberal professions	0.14	0.13
Civil services	0.06	0.05
Industry		
Manufacturing, agriculture, mining and quarrying, construction	0.31	0.33
Trade	0.23	0.23
Services I	0.13	0.13
Services II	0.15	0.14
Public administration, education and health	0.17	0.16
N	**1997**	**2986**

Source: BIBB-CBS 2007.

Textbox A1

Merging question (Excerpt from the CBS 2007 Questionnaire)

We talked a lot about costs and benefits of firm training. To not further extend the interview time, we would like to include data extractions available at the IAB in Nuremburg for the analysis of the survey. These include, for example, information from the Employment History (BeH) data base.

In order to merge these data with the survey data, the law on data protection requires your agreement, which we herewith kindly ask you for. When analyzing this information, it is absolutely ensured, that all data protection requirements are strictly met.

Your agreement to the data merge is voluntary. You can withdraw your agreement at any time.

Do you agree that this additional data information is merged with your interview replies?

1: Yes, agree with data merge

2: No, do not agree with data merge

Source: BIBB CBS 2007 Questionnaire