Schmollers Jahrbuch 132 (2012), 141 – 146
Duncker & Humblot, Berlin

# ALWA-ADIAB – Linked Individual Survey and Administrative Data for Substantive and Methodological Research

By Manfred Antoni and Stefan Seth*

## 1. Introduction

In this paper we present a new data set, which combines two usually distinct data sources: "ALWA survey data linked to administrative data of the IAB" (ALWA-ADIAB). It includes longitudinal and cross-sectional data gathered in the course of the study "Working and Learning in a Changing World" (ALWA)[1] as well as administrative data on the person and the firm level. ALWA-ADIAB opens several new avenues for research, both in substantive and in methodological topics. The data set is made accessible to the scientific community by the Research Data Centre (FDZ) of the Federal Employment Agency at the Institute for Employment Research (IAB), an established provider of German administrative and survey data for the social sciences.

In the fields of educational and labour research, ALWA-ADIAB offers the unique chance of combining the advantages of both survey and administrative data in comprehensive empirical analyses. The ALWA survey provides a wealth of longitudinal and cross-sectional information on its participants' life courses and socio-demographic backgrounds. A particularly important aspect in ALWA is the comprehensive longitudinal measurement of formal and non-formal educational activities. Administrative data, on the other hand, offer accurate longitudinal income and employment histories detailed to the day. Employment spells are supplemented by detailed firm information. Furthermore, the gathering of these data is not subject to non-response, non-compliance or recall error.

ALWA-ADIAB is also suitable for methodological research, such as the validation of longitudinal information given in either of the data sources. As some information is included in both data sources, researchers might also ask

* We would like to thank Iris Dieterich and Peter Jacobebbinghaus for helpful comments.

[1] The acronym is derived from the study's German name "Arbeiten und Lernen im Wandel".

whether the results of empirical analyses depend on the kind of data that is used. Survey researchers may investigate whether certain survey information might be gathered equally reliable from administrative data, in which case data linkage would render such questions redundant in future surveys.

In chapter 2 we give an overview[2] of ALWA-ADIAB by describing the building blocks that constitute the data and the methods used for their linkage. In chapters 3 and 4 we briefly describe the way the data can be accessed and give an outlook on future developments of ALWA-ADIAB.

## 2. The Data

### 2.1 The ALWA Survey

The ALWA data include a wide range of longitudinal and cross-sectional information, gathered in more than 10,400 computerised telephone interviews from August 2007 to April 2008. The sample was drawn from the universe of German residents belonging to the birth cohorts of 1956 – 1988, regardless of their nationality. The survey data contain detailed longitudinal information about complete education and employment trajectories as well as residential, partnership and parenthood histories. Cross-sectional variables inform about, for instance, cultural capital, native language, foreign language proficiency, religiousness and parental background of the respondents. Kleinert et al. (2011) offer an overview of the ALWA data, whereas Antoni et al. (2010) and Matthes / Trahms (2010) provide more detailed information.

### 2.2 The Administrative Data

The administrative part of ALWA-ADIAB consists of comprehensive data on the employment histories of ALWA respondents (individual data), supplemented by data on the establishments they work at (establishment data).

The individual data[3] are provided as spell data. They feature information on employment, unemployment benefits, and job search spells and are accurate to the day, i.e. the data contain exact start and end dates. However, there is no information on spells of self-employment, civil service, and the like. Those appear as gaps in the data. The data on employment and unemployment benefits are available as far back as 1975, while reliable information on job search is available since 2000.

---

[2] Antoni et al. (2011) describe ALWA-ADIAB in full detail.

[3] The ALWA-ADIAB administrative individual data are closely related to the Sample of Integrated Labour Market Biographies (SIAB) data set, described in Dorner et al. (2010).

The following major groups of variables can be found in the administrative individual data:

– Personal characteristics, such as sex, age, school education and vocational training, family background, citizenship.

– Employment status: occupation, gross daily pay, type of job, or daily benefit rate, type of benefit, respectively.

– Job search characteristics, e.g., health status, the willingness to change residence, desired working hours.

– Regional information: place of residence, place of work.

The administrative data originate from two sources: the data on employment stem from a compulsory notification scheme, which requires employers to report on their employees on a yearly basis. The data on unemployment benefits and job search are a by-product of the German Federal Employment Agency's activities, namely job placement and payment of unemployment benefits.

Generally, the administrative data may be considered very reliable. This is particularly true for information that is collected for other than statistical purposes; for instance, the data on remuneration are used by the German Pension Insurance Agency to calculate pension claims. Similarly, the reliability of the information sometimes depends on the data source; for instance, information on education is solid with regard to job search as it is highly relevant for job placement. In contrast, some employers seem not to spend too much effort on reporting correct education information about their employees, which is why this information is often missing or false for employment spells.[4]

The establishment data[5] provide information on the respondents' employing firms on a yearly basis. They are meant to be merged to the individual data, thus enriching them substantially. There is a "basis file" that comprises some basic variables: industrial classification codes, dates of birth and death of the establishment, number of employees, number of marginal part-time workers, median daily wage, and location of the establishment.

On request, "extension files" are available, which offer more detailed establishment data. These include a wealth of information on the establishments' gender and age structures, the distribution of wages within the establishment, and on the qualification structures by various classifications.

---

4 Fitzenberger et al. (2006) propose methods to improve the quality of the education variable. However, their approach relies solely on information from employers. Making use of ALWA-ADIAB survey information allows researchers to broaden this approach substantially.

5 ALWA-ADIAB establishment data are an extract of the Establishment History Panel, described in Hethey-Maier / Seth (2010).

An establishment flow data set is also available on request. These data contain information on gross worker in- and outflows, among others differentiated by gender and age groups. This way it is possible to include indicators of business dynamics into the analysis.

### 2.3 Data Linkage

ALWA-ADIAB is based on ALWA participants, although not all of them are included in the linked data. The requirements for the inclusion were, firstly, that a respondent provided informed consent to the linkage of her survey responses with administrative data; secondly, these consenters had to be identified in the administrative records of the Federal Employment Agency. As the sample was drawn from registers of the residents' registration offices of German municipalities, no unique identifier, such as the social security number, was available in both data sets. Instead, linkage was conducted based on the respondents' names, sex, birth dates and addresses as identifiers, which were compared to addresses drawn from administrative records. Fault-tolerant record linkage techniques were applied to maximise the number of linked respondents (cf. Herzog et al., 2007). The subsequent steps of the record linkage procedure that resulted in ALWA-ADIAB are described by Antoni (2012b). Table 1 presents the number of linked ALWA respondents and the number of observations given for them in the administrative data.

*Table 1*

**Number of Observations in ALWA-ADIAB**

| | |
|---|---|
| ALWA CATI respondents | 10,404 |
| Respondents consenting to data linkage | 9,531 |
| Respondents with linked administrative data | 8,166 |
| Related observations in administrative data | 261,857 |

*Source:* ALWA-ADIAB, own calculations.

Antoni (2012a) examines the selectivity of the linked data and its structure compared to the overall ALWA sample. He finds that participants below 35 and those in dependent employment or registered unemployment are overrepresented in ALWA-ADIAB. By identifying the influence of interviewer and respondent characteristics on the linkage probability, the selectivity of ALWA-ADIAB can be controlled for in multivariate analyses.

ALWA-ADIAB includes a file containing linkage-related information indicating whether a respondent could be linked successfully and if so, how the linkage was achieved and what the quality of the probabilistic match was.

These variables are mainly intended for methodological analyses, which is why they are given for all ALWA participants rather than only for those with a successful link to administrative data.

## 3. Data Access

The legal background of providing access to ALWA-ADIAB is § 75 book X of the German Social Code. ALWA-ADIAB represents so-called social data (*Sozialdaten*), which are subject to strict confidentiality rules. The FDZ has implemented procedures that enable it to comply with data protection legislation while offering data that is altered as little as possible.

The FDZ offers access to ALWA-ADIAB via on-site use with subsequent remote execution. On-site use means that the researcher is granted direct access to the data during a research stay at the FDZ in Nuremberg[6]. After an initial stay, FDZ data users may choose to continue their analyses by sending in scripts (Stata, SPSS, Gauss, TSP) to the FDZ where staff runs the code and returns the results by e-mail. Of course, further research stays are also possible. In either case, output files are checked for compliance with data confidentiality rules and information suited to identify individuals is suppressed.

Prior to data access an application has to be filed by the researcher and be approved by the Federal Ministry of Labour and Social Affairs, and a contract has to be signed. For the application to be successful it must outline a scientific research project concerning social security, which is in the public interest. The FDZ coordinates the application process, which normally takes less than two weeks. Forms and guidance can be found on the FDZ's website[7].

## 4. Outlook: More Administrative Data Sources and Additional Panel Waves

In future versions, ALWA-ADIAB will be extended in several ways. To start with, additional administrative data sources will be included. In particular, future versions will feature information on participation in active labour market programmes (for instance, further training or promotion of employment). Also, based on data usage and user requests, further variables can be added to the data. All FDZ data sets are updated on a regular basis so that there will be more recent observations and an extended observation period.

Finally, the continuation of the ALWA survey as a part of the National Educational Panel Study (NEPS, cf. Allmendinger et al., 2011) will enable re-

6 It is also possible to visit FDZ partner institutions in Berlin, Bremen, Dresden, Düsseldorf, and Ann Arbor, Michigan.

7 http://fdz.iab.de/

searchers to combine the linked respondents from both surveys, though not necessarily under the designation ALWA-ADIAB. This will provide additional observations of survey data for those respondents already included in ALWA-ADIAB, given that they also participated in the NEPS. By adding newly interviewed NEPS participants with a broader age-range than in ALWA, the number of observations increases even more. As the questionnaire of NEPS includes topics that go beyond what has been surveyed during ALWA, the variety of variables and thereby the potential for research will increase even further.

## References

*Allmendinger,* J. / *Kleinert,* C. / *Antoni,* M. / *Drasch,* K. / *Janik,* F. / *Leuze,* K. / *Matthes,* B. / *Pollak,* R. / *Ruland,* M. (2011): Adult Education and Lifelong Learning. In: Blossfeld, H.-P. / Roßbach, H.-G. / von Maurice, J. (eds) (2011): Education as a Lifelong Process. The German National Educational Panel Study (NEPS), Zeitschrift für Erziehungswissenschaft, Special Issue 14, 283 – 299.

*Antoni,* M. (2012a): Linking survey data with administrative employment data: The case of the ALWA Survey, IAB Discussion Paper, forthcoming.

*Antoni,* M. (2012b): Record linkage of the ALWA survey with administrative data of the Federal Employment Agency (ALWA-ADIAB): technical report, FDZ Methodenreport, forthcoming.

*Antoni,* M. / *Drasch,* K. / *Kleinert,* C. / *Matthes,* B. / *Ruland,* M. / *Trahms,* A. (2010): Working and learning in a changing world, Part I: Overview of the study, FDZ Methodenreport 05 / 2010 (en).

*Antoni,* M. / *Jacobebbinghaus,* P. / *Seth,* S. (2011): ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB) 1975 – 2009, FDZ Methodenreport 05 / 2011.

*Dorner,* M. / *Heining,* J. / *Jacobebbinghaus,* P. / *Seth,* S. (2010): The Sample of Integrated Labour Market Biographies, Schmollers Jahrbuch / Journal of Applied Social Science Studies 130 (4), 599 – 608.

*Fitzenberger,* B. / *Osikominu,* A. / *Völter,* R. (2006): Imputation Rules to Improve the Education Variable in the IAB Employment Subsample, Schmollers Jahrbuch / Journal of Applied Social Science Studies 126 (3), 405 – 436.

*Hethey-Maier,* T. / *Seth,* S. (2010): The Establishment History Panel (BHP) 1975 – 2008, FDZ Datenreport 04 / 2010 (en).

*Herzog,* T. N. / *Scheuren,* F. J. / *Winkler,* W. E. (2007): Data quality and record linkage techniques, New York.

*Kleinert,* C. / *Matthes*, B. / *Antoni,* M. / *Drasch,* K. / *Ruland,* M. / *Trahms,* A. (2011): ALWA – New Life Course Data for Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 131 (4), 625 – 634.

*Matthes,* B. / *Trahms,* A. (2010): Arbeiten und Lernen im Wandel. Teil II: Codebuch, FDZ Datenreport 02 / 2010 (de).