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

FDZ-Methodenreport

Methodological aspects of labour market data

03/2018
EN

A Proposed Data Set for Analyzing the Labor Market Trajectories of East Germans around Reunification

Hannah Liepmann,
Dana Müller

Bundesagentur für Arbeit

A Proposed Data Set for Analyzing the Labor Market Trajectories of East Germans around Reunification

Hannah Liepmann (Humboldt-University Berlin) and Dana Müller (Institute for Employment Research – IAB)

Documentation version: DOI: 10.5164/IAB.FDZM.1803.en.v1

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract

Data from German social security notifications and internal procedures of the Federal Employment Agency are an important source for analyzing labor market trajectories. However, for East Germans these data are only fully available from 1992 onwards. As a consequence of German reunification, by 1992 significant fractions of East Germans had already lost their jobs, had changed their occupations and industries, and had moved to West Germany. We partially close the gap in the data by linking the "Integrated Employment Biographies" – that start in 1992 for East Germany – with the GDR's "Data Fund of Societal Work Power" from 1989. The new data set permits the analysis of phenomena such as unemployment, job mobility, and regional mobility. It can also be used to refine the existing knowledge of the individual-level labor market consequences of German reunification. While the GDR add-on is currently not part of our regular data portfolio, our long-term goal is to make the new data set available to the research community via the Research Data Center of the Federal Employment Agency.

Zusammenfassung

Die Daten aus der deutschen Sozialversicherung und den internen Prozessen der Bundesagentur für Arbeit sind eine wichtige Quelle für die Analyse von Arbeitsmarktbiographien. Für Ostdeutsche sind diese Daten allerdings erst ab 1992 vollständig verfügbar. Als Folge der deutschen Wiedervereinigung hatten bis 1992 bereits große Anteile von Ostdeutschen ihre Arbeit verloren, ihre Berufe und Industriezweige gewechselt und waren nach Westdeutschland umgezogen. Wir schließen die Lücke in den Daten teilweise, indem wir die „Integrierten Erwerbsbiographien" – welche 1992 für Ostdeutschland beginnen – mit dem „Datenspeicher Gesellschaftliches Arbeitsvermögen" der DDR aus dem Jahr 1989 verknüpfen. Der neue Datensatz ermöglicht die Analyse von Phänomenen wie Arbeitslosigkeit, berufliche Mobilität und regionale Mobilität. Er kann außerdem genutzt werden, um das bestehende Wissen über die Konsequenzen der deutschen Wiedervereinigung für individuelle Arbeitsmarktbiographien zu verfeinern. Bislang gehört der verknüpfte Datensatz nicht zu unserem regulären Datenangebot. Daher ist unser langfristiges Ziel, den neuen Datensatz für die Forschungsgemeinschaft über das Forschungsdatenzentrum der Bundesagentur für Arbeit verfügbar zu machen.

Keywords: East Germany, German reunification, labor market trajectories, administrative data, record linkage

We are grateful to Chris Bertold of the Federal Archive of Germany for supporting this project and for making it possible. We thank Manfred Antoni for intense discussions, Florian Zimmermann for excellent research assistance, and Niko de Silva and Matthias Umkehrer for valuable suggestions. Financial support by the German Research Foundation (DFG) through the CRC/TRR 190, as well as the support of Bernd Fitzenberger and Alexandra Spitz-Oener, principal investigators of the project A07 of the CRC/TRR 190, is gratefully acknowledged.

1 Introduction

The fall of the Berlin Wall in 1989 and the reunification of Germany in 1990 fundamentally and permanently changed the lives of around 16 Million East Germans (Huinink and Mayer 1995). These unanticipated events overturned the political, economic, and social systems of the former German Democratic Republic (GDR) at a rapid pace. For East Germans, this resulted in new freedoms and opportunities. At the same time, East Germans had to adapt to the new systemic order and were forced to cope with the economic crisis that was caused by the introduction of the market economy (Akerlof et al. 1991, Burda and Hunt 2001).

In the GDR, labor market trajectories were remarkably stable up until 1989. Full employment was guaranteed by the state and GDR citizens had the right, as well as the duty, to work (Grünert 1996, Ritter 2007). After 1989, prolonged unemployment in East Germany meant that drastic job changes (Diewald et al. 1995, p. 322 et seq.) and regional mobility (Hunt 2006, Fuchs-Schündeln and Schündeln 2009) became the norm, rather than the exception. Even today, differences between the former East and the former West remain an important dimension of the persistent socio-economic disparities in Germany. Therefore, thorough and transparent scientific investigations of the individual-level labor market consequences of German reunification are still important today. This need for further research has, for example, been documented in a recent study about perceptions concerning the privatization of East German firms after 1989 (Goschler and Böick 2017).

Administrative data, in particular those stemming from social security notifications and internal procedures of the Federal Employment Agency, are invaluable for analyzing the labor market trajectories of East Germans around reunification. These data are processed at the IAB into a biographical dataset, the so-called “Integrated Employment Biographies” (IEB data), which has a panel structure and large sample size. Moreover, these data provide highly reliable information on a number of key variables, such as average daily wages as well as types and durations of labor market episodes.

However, for East Germans data from social security records are only fully available from 1992 onwards. This is due to the fact that the East German labor market administration was integrated into the West German administration, as part of a complex process. It took time before all firms in East Germany started to report to the social security system (Schmid and Oschmiansky 2007). The resulting gap in the data poses a key empirical challenge. After 1989, significant fractions of East Germans lost their jobs, changed their occupations and industries, or moved to West Germany. A large number of firms closed (e.g., Diewald et al. 1995, Burda and Hunt 2001, Hunt 2006). For many research questions, 1992 is thus too late in time as a starting point for analysis.

Our project “Labor Market Trajectories of East Germans around Reunification” partially closes the gap in the data. For this purpose, we obtained the so-called “Data Fund of Societal Work Power” (in German *Datenspeicher Gesellschaftliches Arbeitsvermögen*, which we abbreviate by GAV data) from the Federal Archive of Germany. The GAV data are a cross-section that provides information on labor market relevant characteristics of around 7 million persons for

the year of 1989. This amounts to 72 percent of the East German labor force at that time. Based on names, exact dates of birth, and gender, we merged the 1989 data with data from social security records that start in 1992. We thus created a unique and very promising new data set that has two major advantages. First, it allows researchers to study mechanisms behind phenomena of general relevance, such as unemployment, occupational mobility, mobility across industries, and regional mobility. Second, it permits the analysis of East German labor market trajectories around reunification based on a sample size that is considerably larger than currently existing data sources. From a historical perspective, the new data therefore enhance the analysis of German reunification. From a political perspective, the new data help refine our knowledge of the causes and consequences of the socio-economic disparities between East and West Germany. These disparities constitute a major dimension of inequality that is still extremely relevant in Germany. Note, however, that at this time, the linked data may only be used within the project “Labor Market Trajectories of East Germans around Reunification.”

This report is structured as follows. In the next section, we present details on the two original data sources that we merged. In Section 3, we describe the merging procedure and evaluate its quality. Section 4 concludes and summarizes the next steps.

2 Original Data Sources

2.1 GAV Data

The so-called “Data Fund of Societal Work Power” constitutes our data source from GDR times. Its German name is *Datenspeicher Gesellschaftliches Arbeitsvermögen*, which we abbreviate by GAV data. “Societal Work Power” is derived from Marxist thought. The authorities in the GDR wanted to refer to a society’s combined knowledge, abilities, and skills that are relevant to economic production (Salomon 1981).

The GAV data were collected in a decentralized way. At the firm and establishment level, human resource departments were required to report information on the characteristics of all employees and had to update these data on a monthly basis. From this source, the councils of each of the fifteen districts of the GDR obtained and combined the information relevant for the GAV data. The councils then transferred this information to the government agency for labor and wages (*Staatssekretariat für Arbeit und Löhne*), which was ultimately responsible for the collection of the GAV data (Gebauer et al. 2004). The quality of these data meets high standards. In particular, the information reported by establishments was fact checked and had to be revised when implausible (Rathje 1996), though in a few instances this revision did not take place and thus resulted in missing information (Dietz and Rudolph 1990).

Neither the original GAV data nor analytical results based on these data were publicly made available. Instead, government agencies in the GDR relied on the GAV data as part of the process of central planning. For example, the data were used to identify and recruit experts demanded in specific circumstances. However, the full potential of the data for central planning purposes was never exploited (Gebauer et al. 2004).

Around 7 million persons are included in the GAV data. Specifically, the data cover the following groups (Dietz and Rudolph 1990, Rathje 1996, Gebauer et al. 2004):

- Workers and employees with a permanent or temporary work contract
- Members of producers' cooperative societies (*Produktionsgenossenschaften*) and law firms (*Rechtsanwaltskollegien*)
- Retired persons still working
- Men performing compulsory military service or alternative civilian service

As is typically the case with GDR official statistics, the GAV data exclude the so-called "Sector X," which was an integral part of the GDR regime. For these employees, separate databases existed. Specifically, the following groups are excluded from the GAV data (ibid):

- Persons working for the Ministry of the Interior, the Ministry of State Security, the Socialist Unity Party, the army, or customs authorities ("Sector X")

Separate databases also existed for specific subgroups, such as certain types of teachers and child care workers. Therefore, some groups are only partially included in the GAV data (ibid):

- The data exclude teachers in schools and child care workers; but include teachers at vocational schools, professors at universities, and employees in nurseries.
- The data exclude the self-employed and their employees; though the majority of craftsmen were members of producers' cooperative societies and are therefore included in the data.
- The data include apprentices; but only those who started apprenticeship training in the year before December 1989.
- The data exclude foreigners temporarily working in the GDR under the coverage of intergovernmental agreements; but include foreign GDR residents.

For the workers who are included in the GAV data, rich information was elicited. The variables can be divided into four categories:

1. Demographic characteristics include age, gender, place of residence, the number of children under 14 and the number of persons in need of care in the household, disability status, marital status, and nationality.
2. Qualification characteristics include high school education, current apprenticeship training, and university degree.
3. Employment characteristics include the type of employment, place of employment, leave of absence, main job task, job status, work hours, and occupation.
4. Firm characteristics include firm type and industry.

In our project, we use a cross-section of the GAV data that refers to December 31 in 1989. The history of these data demonstrates that their survival was not self-evident: In fact, the GAV data had been collected on an annual basis[1]. However, due to limited computer capacities in the GDR, only data from the current year were kept while data from previous years were deleted (Gebauer et al. 2004). The Federal Archive of Germany obtained the 1989 GAV data in November of 1991 on magnetic tape. In 1998, the data were for the first time saved on CD-ROM. In addition to the GAV data, detailed wage information had been collected for 2.3 Million employees in the GDR. Unfortunately, because of data protection regulations, these wage data were deleted in 1991 in reunified Germany (Rathje 1999). The data that we use in our project therefore do not contain any information on pre-unification wages.

The vast majority of historians and other scientists conducting research in the Federal Archive of Germany do not employ quantitative methods (Rathje and Wettengel 1999). Therefore, up until today, few researchers have analyzed the GAV data. The studies that we are aware of include Salomon (1981) and Groebel (1997). Salomon (1981) provides a technical report concerning the processing and analysis of the GAV data. This report reflects the information technology available in the GDR in the early 1980s[2]. Groebel (1997) explores reasons for the divergence of sectoral employment structures in market economies and planned economies. Among other data sources Groebel (1997) relies on the GAV data to provide descriptive statistics that illustrate her theoretical arguments. Additionally, from 2001 through 2012, a subproject of the Collaborative Research Center 580 studied GDR elites and used the GAV data, though only as a supplementary source of information (Gebauer et al. 2004, Salheiser 2006).

2.2 IEB Data

We merged the GAV data with data from the so-called “Integrated Employment Biographies” (IEB data). The IEB data are a natural choice for our project, because they contain labor market relevant information that resembles the GAV data. The IEB data include information from two sources: social security notifications and internal processes of the Federal Employment Agency. We discuss each of these sources in turn.

First, social security notifications involve an integrated notification procedure for the health, pension, and unemployment insurance programs, which is known by the abbreviation DEÜV (for more details see Wermter and Cramer 1988, Bender et al. 1996, p. 4 et seq.). It has been mandatory since 1973 in West Germany and since 1991 in East Germany. The notifications include several pieces of information on all insurable employment episodes reported by every employer. Section 28 of the Social Code Act IV determines what kind of information needs to be notified. In general, a notification includes information about the beginning and end of each employment episode that is subject to social security contributions, as well as corresponding information about gross wages, education, employment status, occupation and nationality. In

[1] The GAV data project was initiated in 1975, though it took until 1986 before the data were made fully available to the government agencies interested in them (Gebauer et al. 2004).
[2] Salomon (1981) constitutes a dissertation written at Humboldt-University Berlin (East). During GDR times, this dissertation was classified as confidential. After reunification, a copy was retrieved by the Federal Archive of Germany such that the dissertation can now be accessed by the public.

addition, there is a mandatory notification for every employer liable to social security contributions at least once a year. Since 1999 employment episodes of marginal part-time employees and family workers have also been recorded. Importantly, the social security notifications do not include civil servants, self-employed individuals and regular students.

The following will give a more detailed idea of the notification procedure. The data are recorded by the health insurance companies first, and then are transmitted to the German pension insurance, which in turn forwards the data to the Federal Employment Agency. The data are collected and processed by the Federal Employment Agency, particularly for generating employment statistics. Subsequently the data are processed into employment histories at the Data and IT-Management Department of the IAB. These employment histories constitute the so-called Employee History File, which starts in 1975 for West Germany and includes East Germans from 1992 onwards.

Second, internal processes of the Federal Employment Agency are the other source of the IEB data. These data are collected to fulfill legal requirements, to inform the public and in the preparation of statistics. The data are then prepared at the IAB and organized in four different history files:

1. The Benefit Recipients History includes all periods during which unemployed individuals received earnings replacement benefits from the Federal Employment Agency within the scope of Social Code Book III (SGB III). The data start in 1975.

2. The Unemployment Benefit II Recipient History covers all periods during which unemployed individuals received benefits in accordance with the Social Code Book II (SGB II). It was implemented in 2005 and captures the pooling of unemployment benefits and social assistance. The difference compared with SGB III is that unemployment benefits are not determined individually but depend on the so-called “benefit community” (which includes certain household members, such as spouses and children). This data source only contains information about individuals who are capable of working or are under the age of 64, and about the benefit community’s members in accordance with Section 7 of SGB II. However, the Federal Employment Agency is not the only responsible authority for administering the benefits that fall under Social Code Book II. There are three possible types of institutions the data can stem from:

    a. Joint facilities of employment agencies and municipalities since 2011 (before 2011 cooperation of employment agencies and municipalities in the context of so-called working partnerships);

    b. separated responsibilities until 2011 with divided tasks between the Federal Employment Agency and the municipality; and,

    c. authorized municipalities which are responsible for all tasks regarding the SGB II.

The data originate from different reporting procedures. In particular, authorized municipalities can use their own IT procedures and transmit the data to the Federal Employment Agency. The data have been collected since 2005 but the data are complete only from 2007 onwards.

3. The Participation-in-Measure History Files include active labor market policy measures within the scope of SGB III and in accordance with SGB II if these measures are reported in Federal Employment Agency IT procedures. The data are available from 2000 onwards.

4. The Jobseeker History contains information on jobseekers who are registered with employment agencies. The data are available from 2000 onwards and were expanded in 2005 to also include jobseekers receiving Unemployment Benefit II.

Finally, the data from the social security notifications (i.e., the Employment History File) and the data from the internal processes of the Federal Employment Agency (i.e., the four other history files just described) are combined. Together, these data sources represent the Integrated Employment Biographies (IEB data).

Note that the IEB data could not directly be used for the linkage procedure with the GAV data, because the IEB data lack direct identifiers like names for reasons of data privacy. As we explain in more detail in the next section, we instead used information from the data ware house of the Statistics Department of the Federal Employment Agency. This information includes all individuals from the IEB data as well as their direct identifiers. Variables from the IEB data will then later be merged to the linked new data set.

3 Newly Created Data Source

3.1 Procedure for the Merge of the GAV Data and the IEB Data

For the purposes of merging the GAV data and the IEB data, the Federal Archive of Germany provided us with the non-anonymized version of the 1989 GAV data. We received fifteen Excel documents, each referring to one of the districts in the GDR, which we transformed into a single file in Stata format. Based on the non-anonymized version of the GAV data, we were able to exploit the following information for the merge: first name, last name, exact date of birth, and gender. In principle, it would have been possible to rely on additional information pertaining to occupations, industries, and regions. However, the IEB data are fully available for East Germans only from 1992 onwards. Between 1989 and 1992, a significant fraction of East Germans changed jobs and moved between regions. Hence, using this additional information would have led to oversampling of persons who did not move across regions or did not change jobs. In order to avoid such biases, we deliberately decided not to use the additional information for the merge and relied on names, date of birth, and gender only.

From the data warehouse of the Statistics Department of the Federal Employment Agency we similarly obtained information on names, date of birth, and gender for persons covered by the IEB data. In addition, we obtained their anonymized personal IDs that will later allow us to merge further IEB variables. Our aim was to identify workers from the GDR and to reduce the complexity of the data. Therefore, when drawing from the universe of individuals included in the IEB data we imposed three restrictions. First, we focused on persons born between 1929 and 1976 who were aged 13 to 70 in 1989. Second, we only included persons for whom at least one episode is recorded in the IEB data between 1990 and 1996 in East or West Germany. Third, we imposed that for these persons no such episodes were recorded in West Germany before 1990. Because of the third criterion, a large number of West Germans are excluded from the merging procedure. This reduces the likelihood of false matches. At the same time, it implies that we neglect individuals who migrated from West to East Germany before the wall fell. However, only few West Germans moved to the GDR during this period (see for example the graph in Hunt 2006, p. 1017). Note that, for the individuals who we match, selected variables on their entire history from the IEB records will ultimately be included in the data set.

We conducted the merge in collaboration with Manfred Antoni who describes the technical details of the procedure in Antoni (2018). We began by preprocessing the GAV data. Duplicates in the GAV data were one issue we needed to address (see Table 1). On the one hand, this concerns pure duplicates, where all variables are identical to an original observation. We dealt with these cases by dropping all 166,604 pure duplicates. On the other hand, there are cases of data entries that contain information on multiple jobs held by the same individual. Specifically, there are 194,916 data entries which refer to individuals' second or higher order observations (Table 1). We do not know with certainty whether these are observations referring to individuals performing several jobs in parallel or observations referring to individuals' previous jobs. The latter case would refer to situations in which the data were not updated after job changes. We were, however, able to code a variable ordering multiple data entries per person by the date the data were collected. One possibility is therefore to restrict the analysis to each individual's most recent job spell, which we did when merging[3]. This left us with a sample size of more than 7 million persons included in the GAV data (Table 1).

[3] When we performed the merge, we also encountered the issue of duplicates in the IEB data. These are cases where we found matches based on exact names and date of birth but where we could not identify a unique IEB person ID. However, this concerns few cases (< 1%) which we treat as unsuccessful matches.

Table 1 Number of Observations and Persons in the GAV Data and Merging Quotas

Original number of observations in GAV data	7,412,001
*Among these: Number of pure duplicates**	166,604
Actual number of observations in GAV data	7,245,397
*Among these: Second or higher order observations for persons in the GAV data***	194,916
Number of persons in GAV data	7,050,481
Among these: Persons with name of four letters or less	16,406
Number of persons in GAV data merge was based on	6,978,591
Among these: Last name and first name available	6,479,700
Among these: Last name available only or first name and last name not separated by a comma	498,883
Number of GAV persons identified in IEB data	5,407,817
Percentage of persons in GAV data for whom a match was found	
All	0.7670
Women Only	0.7240
Men Only	0.8048
Younger than 60, all	0.8221
Younger than 60, women only	0.7680
Younger than 60, men only	0.8706

* All variables were identical

** Second or higher order entry for a person with identical name, date of birth, and gender

We next preprocessed the information on names. In the original GAV data, information on names is presented in the format of "*last_name, first_name*". While this is true for the majority of names, we had to account for the fact that some names deviated from the intended format. To ensure comparability across data sources, identical preprocessing steps were applied to both the GAV data and the IEB data. In particular, the following steps were necessary:

- Special characters, which appeared in various formats and at varying places in the name variable, were deleted or replaced by a comma to separate first and last names (for example when the name was in the format of "*,,last_name, first_name"* or *"last_name. first_name"* etc.)

- Name suffixes, which appeared in various formats and at varying places in the name variable, were deleted. This was relevant for academic titles (such as "Dr." or "Profes-

sor"), titles of the nobility (such as "Von", "Graf" or "Freiherrin"), and generational designations (such as "Junior" or "Sr."). It also concerned farmers with a supplementary last name (such as "last_name1 genannt last_name2" where the suffix "genannt" was deleted).

In the GAV data, additional peculiarities in the name variable required further investigation. We therefore made the following adjustments:

- There are cases in the GAV data where the information for "*last_name, first_name*" consists of less than four letters. Often, these letters do not constitute plausible names. In some cases, these letters refer to actual, short last names, where no information on the first name is provided. According to experience from previous data record linkages performed at the Research Data Center of the Federal Employment Agency, it is highly unlikely to find matches in the IEB data based on name information of four letters or less. Therefore, we decided to exclude these cases from the merge. This concerned around 16,000 persons (see Table 1).

- In the GAV data, for around 500,000 persons, the information "*last_name, first_name*" consists of one word only. In the majority of cases, one-word-names refer to last names. We therefore interpreted these names as last names and used this information for the merge. Additionally, there are one-word-names including a last name and a first name which lack a separating comma. In order to distinguish between first and last names, we used a routine that identified and separated common first names. We then fact checked these results manually, since additional corrections were required that could not be automated (for example when a common first name was in fact part of a last name as in "Franke" ,"Schubert" or in more exotic semantic combinations).

After preprocessing, we based the merge on 6.98 million persons included in the GAV data. For 93 percent of these, we used information of first and last names, whereas for the remaining 7 percent either the last name was available only or first and last names were not separated (Table 1). To put these figures into perspective, the GDR labor force of 1989 consisted of 9.75 million persons (Federal Statistical Office 1994). Thus, our merge encompasses 72 percent of the East German labor force.

3.2 Evaluation of the Merging Procedure

For 77 percent of persons from the GAV data we found a match in the IEB data (Table 1). According to experience from previous merges performed at the Research Data Center of the Federal Employment Agency, this is a good quota.

For the vast majority of matches (88 percent), the information on first and last names, date of birth, and gender was identical in both the GAV and the IEB data (data not shown). The remaining fraction was matched using record linkage techniques that tolerate a justifiable degree of error while at the same time keeping the likelihood of false matches as small as possible. Three steps were particularly relevant in increasing the merge quota. First, we tolerated small spelling or coding mistakes in the name information and in the day of birth, but imposed that

the other identifiers (gender, month and year of birth) were matched accurately. Second, we required a perfect match between last names, birth date, and gender but dropped first names. This step was especially important; as a significant fraction of persons in the GAV data lack a first name (see Table 1). Third, we repeated the previous step but relied on first names while neglecting last names. We only kept cases where a unique match was found. Manfred Antoni performed these steps and provides more details on the exact implementation in Antoni (2018).

We next use OLS-regression analysis to investigate how the success of the merge correlates with key observable characteristics. Specifically, we regress a dummy variable that is equal to one in case of a successful merge on key observable characteristics in 1989. Key observable characteristics are measured as categorical variables and refer to gender, age intervals, type of school diploma obtained, and marital status. The regression results are displayed in Table 2. For reference, we display corresponding summary statistics for the independent variables in Table 3.

Based on Column (1) in Table 2, three phenomena should be emphasized. First, the merge quota is considerably lower for individuals older than 60 in 1989. Indeed, we were not able to find any matches above the age threshold of 61 (data not shown). This is due to the fact that for the IEB data the information on names stems from the late 1990s. By this time, older East Germans had dropped out of employment. Most analyses based on the new data set should therefore be limited to persons younger than 60. If these are dropped, the quota of successful merges increases to 82 percent (see Table 1). Below the age threshold of 50 in 1989, the same quota increases further to 86 percent (data not shown). In our view, these are high merge quotas that speak for the quality of the new data set. At the same time, these quotas highlight that the data are more reliable for persons younger than 60 or even 50 in 1989.

Second and reassuringly, there are only negligible differences by qualification level in the success of the merge. This can be seen in Column (1) of Table 2, where we measure qualification in terms of four different levels of school diploma obtained. We used these categories because the usual distinction between an apprenticeship degree and no formal vocational qualification cannot be made in the GAV data. The only noteworthy difference by qualification level is the significantly lower merge quota for persons with missing information about the school diploma obtained. However, this concerns very few cases (< 0.5 %, see Table 3).

Third, the merge quota is considerably lower for women than for men. For women younger than 60 in 1989, this quota amounts to 77 percent, which is around 10 percentage points lower than the merge quota of their male counterparts (see Table 1). We investigate this further in Columns (2) and (3) of Table 2, where we perform regression analysis as before but this time split the sample by gender. Furthermore, we exclude persons older than 59.

Table 2 OLS Regression Results Assessing the Success of the Merge by Gender, Age, Qualification Level, and Marital Status

	(1) All	(2) Women, < 60 years	(3) Men, < 60 years
Female	-0.1003***		
	(0.0003)		
Age intervals			
≥ 20 & < 30 years	0.0047***	0.0064***	-0.0121***
	(0.0008)	(0.0014)	(0.0008)
≥ 30 & < 40 years	0.0351***	0.0744***	-0.0128***
	(0.0008)	(0.0014)	(0.0009)
≥ 40 & < 50 years	0.0371***	0.0975***	-0.0253***
	(0.0009)	(0.0015)	(0.0010)
≥ 50 & < 60 years	-0.1557***	-0.2111***	-0.1112***
	(0.0009)	(0.0016)	(0.0010)
≥ 60 years	-0.8137***		
	(0.0009)		
School Diploma			
8 years of schooling	-0.0101***	-0.0189***	0.0009
	(0.0006)	(0.0011)	(0.0007)
10 years of schooling	0.0014**	0.0001	0.0041***
	(0.0006)	(0.0011)	(0.0007)
Abitur (12 years)	-0.0237***	-0.0211***	-0.0208***
	(0.0007)	(0.0014)	(0.0009)
Missing	-0.1166***	-0.1009***	-0.1695***
	(0.0137)	(0.0217)	(0.0228)
Marital Status			
Married	0.0846***	0.1650***	0.0269***
	(0.0004)	(0.0008)	(0.0005)
Widowed	0.0114***	0.0399***	-0.0290***
	(0.0010)	(0.0019)	(0.0028)
Divorced	-0.0111***	0.0122***	-0.0175***
	(0.0007)	(0.0011)	(0.0009)
Missing	0.0283**	0.0837***	-0.0100
	(0.0143)	(0.0224)	(0.0238)
Constant	0.8355***	0.6659***	0.8914***
	(0.0009)	(0.0016)	(0.0009)
R squared	0.2753	0.1077	0.0153
N	7,050,410	3,102,800	3,453,027

Notes: *OLS regression analysis of a dummy variable equal to 1 in case we were able to find an individual included in the GAV data also in the IEB data and equal to 0 otherwise on key observable characteristics measured in 1989 (gender, age intervals, type of high school diploma obtained, and marital status). Reference categories are: male (Column (1) only), age below 20, no high school diploma indicating at least eight years of schooling, and being married. The sample includes all individuals from the GAV data, but excludes observations when the same individual is included in the GAV data more than once (see Table 1). In column (1), very few individuals are excluded from the regression due to missing age information. Robust standard errors are in parenthesis. *** and ** denote significance at the 1 and 5 percent levels.*

Table 3 Summary Statistics (in Percent)

	(1) All	(2) Women, < 60 years	(3) Men, < 60 years
Female	46.73		
Age intervals			
< 20 years	4.58	4.81	5.02
≥ 20 & < 30 years	23.29	25.95	24.24
≥ 30 & < 40 years	24.77	26.32	26.92
≥ 40 & < 50 years	19.67	21.04	21.25
≥ 50 & < 60 years	20.68	21.88	22.57
≥ 60 years	7.01		
Missing	0.00		
School Diploma			
< 8 years of schooling	6.84	4.92	9.10
8 years of schooling	40.74	37.22	37.55
10 years of schooling	44.38	50.28	44.26
Abitur (12 years)	7.84	7.31	8.94
Missing	0.20	0.26	0.16
Marital Status			
Single	22.34	19.50	27.46
Married	67.18	68.46	64.92
Widowed	2.35	2.46	0.64
Divorced	7.95	9.33	6.83
Missing	0.19	0.25	0.14
N	7,050,481	3,102,800	3,453,027

Notes: Summary statistics for independent variables used in the regression analysis presented in Table 2. All variables refer to 1989.

We show that the likelihood of finding a match among women is lowest for initially single women. In particular, the merge quota is considerably higher for those who are married in 1989 compared with those who are single initially[4]. This indicates that the lower merge quota among women can be rationalized by the fact that initially single women changed their names after marriage and are then more difficult to identify in the IEB data[5]. Our recommendation is therefore that future analyses should include robustness tests concerning women's initial marital

[4] With 68 and 20 percent, respectively, initially married and initially single women are the largest group; 9 percent of women are divorced, and 2 percent are widowed (see Table 3, Column 2).

[5] In general, when East German women married for the first time, they usually adopted the husband's family name, but kept this name after a divorce. Thus, for our merge we are particularly concerned about women's transitions from initially being single to being married, while we are less concerned about transitions from initial marriages into divorce. Our reasoning is based on anecdotal evidence from discus-

status. It might for example be appropriate to add marital status as a control variable or to investigate whether main effects differ by marital status category.

4 Conclusions and Next Steps

This report summarizes our linkage of the GDR's "Data Fund of Societal Work Power" from 1989 with the Integrated Employment Biographies from later years. The merge was based on around 7 million East German workers, which amounts to 72 percent of the East German labor force at that time. We were able to obtain a comparatively high merge quota of 82 percent among persons younger than 60 in 1989. However, the merge was somewhat less successful for older workers, who dropped out of the labor force, and for initially single women, who often changed their names after marriage.

Before the linked data can be used in the project "Labor Market Trajectories of East Germans around Reunification," different processing steps are still necessary. First of all, the direct identifiers such as names have to be deleted and will be replaced by a pseudo-identifier. Variables from the IEB, which are normally contained in the linked standard data products of the Research Data Center (FDZ), such as PASS-ADIAB (see Antoni and Bethmann 2018), will be added to the new data set. In addition, we need to develop crosswalks in order to harmonize variables that are similar in the GAV and IEB data, such as the occupation, industry, and region variables. Finally, a sample needs to be conceptualized and drawn from the linked data.

The linked data will allow the analysis of research questions about East German employment biographies around reunification based on sample sizes that are much larger than those of currently existing data sources with panel structure. However, our legally binding agreement with the Federal Archive of Germany requires that the data will have to be deleted after the end of our project. Therefore, the Research Data Center (FDZ) of the Federal Employment Agency will develop a concept in order to convince the Federal Archive to make the linked data available to the research community via the FDZ. We believe that the new data have considerable potential to answer research questions of general scientific interest, as well as to enhance the understanding of differences between East and West Germany, which still constitute a major dimension of persistent socio-economic disparities in this country.

sions with former employees of the former population registration office in Karl-Marx-Stadt (now Chemnitz). We are not aware of systematic studies that analyze norms concerning female names for East Germany.

References

Akerlof, George, Andrew Rose, Janet Yellen, and Helga Hessenius (1991). “East Germany in from the Cold: The Economic Aftermath of Currency Union.” *Brookings Papers on Economic Activity* 1, 1-105.

Antoni, Manfred and Arne Bethmann (2018). PASS-ADIAB – linked survey and administrative data for research on unemployment and poverty. *Jahrbücher für Nationalökonomie und Statistik*, online first, 10 pages.

Antoni, Manfred (2018). Record linkage of GDR’s “Data Fund of Societal Work Power” with administrative labour market biography data of the German Federal Employment Agency. *German RLC Working Paper No. wp-grlc-2018-02.*

Bender, Stefan, Jürgen Hilzendegen, Götz Rohwer, and Helmut Rudolph (1996). Die IAB-Beschäftigtenstichprobe 1975-1990. *Beiträge zur Arbeitsmarkt- und Berufsforschung* 197, Nuremberg.

Burda, Michael and Jennifer Hunt (2001). “From Reunification to Economic Integration: Productivity and the Labor Market in Eastern Germany.” *Brookings Papers on Economic Activity* 32(2), 1-92.

Dietz, Frido and Helmut Rudolph (1990). Berufstätigenerhebung und der Datenspeicher „Gesellschaftliches Arbeitsvermögen“, *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 4, 511-518.

Diewald, Martin, Johannes Huinink, Heike Solga, and Annemette Sorensen (1995). Umbrüche und Kontinuitäten – Lebensverläufe und die Veränderung von Lebensbedingungen seit 1989. In Huinink, Johannes and Karl Ulrich Mayer (Eds.) *Kollektiv und Eigensinn. Lebensverläufe in der DDR und danach,* pp. 307-348. Berlin: Akademie Verlag.

Federal Statistical Office (1994). Sonderreihe mit Beiträgen für das Gebiet der ehemaligen DDR, Heft 14, Erwerbstätige 1950 bis 1989, Wiesbaden.

Fuchs-Schündeln, Nicola and Matthias Schündeln (2009). Who Stays, Who Goes, Who Returns? East-West Migration within Germany since Reunification. *Economics of Transition*, 17(4), 703-738.

Gebauer, Ronald, Dietmar Remy, and Axel Salheiser (2004). Der Datenspeicher „Gesellschaftliches Arbeitsvermögen“: prozessproduzierte Daten als Quelle für die quantitative historische Sozialforschung und eine Soziologie des DDR-Sozialismus. *Historical Social Research*, 29(4), 196-219.

Goschler Constantin and Marcus Böick (2017). Wahrnehmung und Bewertung der Arbeit der Treuhandanstalt, Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie, http://www.bmwi.de/Redaktion/DE/Publikationen/Studien/wahrnehmung-bewertung-der-arbeit-der-treuhandanstalt-lang.html (accessed January 2018).

Groebel, Annegret (1997). *Strukturelle Entwicklungsmuster in Markt- und Planwirtschaften, Vergleich der sektoralen Erwerbsstrukturen von BRD und DDR.* Heidelberg: Physica Verlag.

Grünert, Holle (1996). Das Beschäftigungssystem der DDR. In: Burkart Lutz, Hildegard Nickel, Rudi Schmidt, and Arndt Sorge (Eds.) *Arbeit, Arbeitsmarkt, und Betriebe*, pp.17-69. Opladen, Berlin, Toronto: Barbara Bulich.

Huinink, Johannes and Karl Ulrich Mayer, eds. (1995). *Kollektiv und Eigensinn. Lebensverläufe in der DDR und danach.* Berlin: Akademie Verlag.

Hunt, Jennifer (2006). “Staunching Emigration from East Germany: Age and the Determinants of Migration.” *Journal of the European Economic Association* 4(5), 1014-1037.

Rathje, Ulf (1996). Der „Datenspeicher Gesellschaftliches Arbeitsvermögen“ der DDR, *Historical Social Research,* 21(2), 113-117.

Rathje, Ulf (1999). Bestand DQ 3 MD, Datenspeicher Gesellschaftliches Arbeitsvermögen, Bundesarchiv Koblenz, 104 pages.

Rathje, Ulf and Michael Wettengel (1999). Digitale Datenbestände von Behörden und Einrichtungen der DDR im Bundesarchiv, *Historical Social Research,* 24(4), 70-101.

Ritter, Gerhard (2007). Rahmenbedingungen der innerdeutschen Einigung. In Gerhard Ritter and Federal Ministry of Labour and Social Affairs and The Federal Archive (Eds.) *Geschichte der Sozialpolitik in Deutschland, Band 11, Bundesrespublik Deutschland 1989-1994*, pp. 343-395. Baden Baden: Nomos Verlag.

Salheiser, Axel (2006). Der Datenspeicher „Gesellschaftliches Arbeitsvermögen“ (DS GAV) des Staatssekretariats für Arbeit und Löhne, in: Best, Heinrich und Dietmar Remy (Eds), Die geplante Gesellschaft, Analysen personenbezogener Datenspeicher der DDR, SFB-580-Mitteilungen, Heft 18, 117-128.

Schmid, Günther and Frank Oschmiansky (2007). Arbeitsmarktpolitik und Arbeitslosenversicherung. In Gerhard Ritter and Federal Ministry of Labour and Social Affairs and The Federal Archive (Eds.), *Geschichte der Sozialpolitik in Deutschland, Band 11, Bundesrepublik Deutschland 1989-1994*, pp. 435-491. Baden Baden: Nomos Verlag.

Solomon, Jürgen (1981). Probleme bei der Genauigkeit bei der Massendatenverarbeitung unter besonderer Berücksichtigung der Fehlerbereinigung und der Fortschreibung, dargestellt am Beispiel des Projektes „Gesellschaftliches Arbeitsvermögen“, PhD Thesis, Humboldt-University Berlin (East).

Wermter, Winfried and Ulrich Cramer (1988). Wie hoch war der Beschäftigtenanstieg seit 1983? Ein Diskussionsbeitrag aus der Sicht der Beschäftigtenstatistik der Bundesanstalt für Arbeit. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung,* 4(88), 468-482.

Imprint

FDZ-Methodenreport 3/2018 (EN)

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

Editorial staff
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2018/MR_03-18_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Hannah Liepmann
Humboldt-University Berlin
Email: hannah_liepmann@gmx.de

Dana Müller
Institute for Employment Research (IAB)
Research Data Centre (FDZ)
Email: Dana.Mueller@iab.de