Projekt
Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland

---

Arbeitspapier Nr. 8 des Projekts Ausbildungs- und Berufsverläufe
der Geburtskohorten 1964 und 1971 in Westdeutschland

Maike Reimer
Ralf Künster

Linking Job Episodes from Retrospective Surveys and Social Security Data: Specific Challenges, Feasibility and Quality of Outcome

08/2004

Contact:

Maike Reimer
Ralf Künster
Max-Planck-Institut für Bildungsforschung
Lentzeallee 94
14195 Berlin

reimer@mpib-berlin.mpg.de
kuenster@mpib-berlin.mpg.de
http://www.mpib-berlin.mpg.de

Contents

1 Introduction 4

2 Matching Data: Theoretical Framework 4

2.1 Developing a Match Rule for Register Data from IAB and Survey Reports from the GLHS - The Two Data Sources 6
2.2 How Job Episodes and Their Start and End Dates Are Generated in the IAB File 8
2.3 Construction of Employment Episodes in the GLHS 9
2.4 The Matching Procedure 11

3 Matching Employment Episodes – Data Preparation and Match Rates 13

3.1 Sample 13
3.2 Employment Episodes in IAB and GLHS 14
3.3 Success of the Matching Procedure 18
3.4 Qualitative Analysis of Match Failures 20

4 Conclusion and Outlook 22

4.1 What Have We Learned About Match Rates and the Causes of Match Failure? 22
4.2 Theoretical and Practical Implications 23

5 Literature 26

6 Appendix 28

Abstract

This paper deals with the linking of longitudinal survey and register data from the German Social Security agencies on the basis of events or episodes. Events, episodes or transitions are products of error-prone data generation processes and therefore can differ between the data sets to be linked. A match rule that uses starting and ending date of each episode as identifying information is developed, applied and evaluated. It distinguishes between *perfect matches*, *time-liberal matches* and *multiple matches*. About 70 % of all episodes can be successfully matched. Qualitative case studies show that linkage failure is due to the differing construction of events/episodes, errors in the recall of the identifying information and in the recall of the preparatory information in the survey reports. Three issues are discussed: 1.How should a specific time period in the linked data file be represented? 2. What is the relationship between manual linkage on the basis of case studies and automatic linkage rules? 3. What are the implications for the validity of the survey data?

1 Introduction

The empirical study of individual employment careers as a sequence of episodes in various labor market contexts is indispensable for sociological, economical and social-scientific analyses of societies and societal change over time. The necessary empirical data most often come from retrospective or panel surveys, conducted by scientific or official institutions such as the census bureau. Another way to get relevant information are the registers that administrative institutions and offices assemble in the process of fulfilling their duties (so called process-produced data). The linkage of two data sources with complementary information about the same objects or units (human beings, organisations, events) will increase explanatory power through the combination of variables. For this purpose, the objects or units in the two data files must be matched correctly, ideally by a computer program that decides on the basis of automatic linkage rules. Most frequently, the units or objects to be linked are individuals; therefore literature on the linking of longitudinal register and survey data on the level of events or episodes is scarce and consequently, little is known about the specific challenges.

In this paper, we will examine this question using two data sets: a) retrospective survey reports about employment episodes from the „German Life History Study" (GLHS), and b) data from the German social security agencies about jobs held by the respondents, compiled by the Institute for Employment Research (IAB) of the Federal Employment Services' (BA, "Bundesanstalt für Arbeit"[1]) of the Federal Republic of Germany. First, we will look into the theoretical and formal framework for the linkage and its evaluation. Second, we will describe both data sets and the process of data generation in detail and make assumptions about the discrepancies that are to be expected. On this basis, we will develop a match rule for employment episodes. Third, we will report the preparatory data transformation and describe the resulting match rates of the matching procedure. Using selected case studies, we will identify reasons for match failure and at last discuss the practical and conceptual implications.

2 Matching Data: Theoretical Framework

Fellegi and Sunter's (1969) classical framework for data linkage assumes two data sets that represent information about the same objects – usually individuals - in a similar matrix format: each object is represented by a row or record and characterized by values of variables in the columns. A human being or a computer program can then match corresponding objects by

[1] From January 2004 „Bundesagentur für Arbeit".

comparing every possible pair of records with respect to one or more selected identifying variables (i.e., name or individuals, birth dates, birth places) and, using a predetermined criteria, making one of three decisions about the classification of two records:

- they do represent the same objects and are finally matched („match"),
- they do not represent the same objects and are finally not matched ("non-matched"),

it is impossible, on the basis of the available information, to decide whether they represent the same objects or not and the match decision is postponed.

Central elements of the matching procedure and crucial for its success are therefore the *selection of identifying information* and the *selection of the rules and criteria*, that decide about the match status.

Identifying information must be available in both data sources in comparable form and for all records. Moreover, it must identify (individually or in combination) uniquely one and only one object or unit. Since data generation processes lead inevitably to errors in the data, records that represent the same object can differ regarding the identifying information, and records that represent different objects can be consistent in their identifying information. Thus, the matching procedure can produce two kinds of misclassifications:

- records that represent the same objects are erroneously classified as non-match ("type-A-error"), and
- records that do not represent the same objects are erroneously classified as match ("type-B-error").

Rules and criteria generally should be chosen in a way that minimizes the risk for errors of Type A an B simultaneously and produces both a high hit rate and a low rate of misses. Unfortunately, match rules that have a lower risk for Type-A-errors tend to have a higher risk for Type-B-errors: a liberal match rule that decides "in favor of the plaintiff" whenever there is room for doubt will more often declare records that represent different objects as matched. The other way round, a conservative and strict rule that rather would declare a doubtful case as non-match will more often overlook records that represent the same objects and differ in their identificatory information only by error. Depending on the purpose of the matching, rules and criteria must be selected more conservatively (specifically) or more liberally (sensitively) to place special emphasis on either avoiding Type-A errors or type B errors.

So far the basic assumptions of Fellegi and Sunter. When matching events or episodes, another circumstance is to be considered. Episodes are temporally extended events or states that are delimited by changes in the relevant defining features (such as activities, roles, places). They are

not "natural units" in the sense that individuals are. They are by their very nature fuzzy, that means, they can be delimited by many different and even by mutable criteria. The definition of episodes in a data set reflects the interests of the data collection agents. For register data, the administrative necessities of the data gathering institution are crucial. For social scientific survey reports, the social scientific concepts regarded are as useful when describing the selected aspects of reality and when answering the researcher's substantial questions. As observed by Luks and Brady (2003, 416), „there is a fundamental identification problem when we use survey data to cross-check administrative data (about episodes of welfare recipiency)". Also Miller and Groves (1985) bring forth the question of the conceptual nature of objects or units under observation in their attempt to link events of victimization in police registers and survey reports.

Therefore, the episodes in survey and register data themselves are outcomes of differing data generation processes and therefore can differ in the two data sources. To illustrate this point, take the following example: a respondent reports employment for all of 1995 and a new job in a different company from January 1996 for the next six months until July 1996. The employers' register however states that the employment this person held from January 1995 on ends already in November of 1995 and the next employment episode starts in February 1996. Such discrepancies can sometimes be explained by assumptions about differences in the two data generation processes. Here, we might argue that the register is correct while the respondent has misdated end and start of these two consecutive jobs and forgotten a short time of unemployment between them. The challenge is to decide whether or under which circumstances such episodes from survey and register can be meaningfully matched and how this can be translated into a computerized linkage rule. To meet this challenge, the processes by which episodes are created in the survey and register data and the kind of discrepancies that occur must be known very well and in much detail.

2.1 Developing a Match Rule for Register Data from IAB and Survey Reports from the GLHS - The Two Data Sources

Our register data is based on the mandatory integrated notification procedure for health insurance, statutory pension schemes and unemployment insurance that was introduced on 1 January 1971 (DEVO/DEVÜ)[2] (Bender and Hilzendegen, 1995). Employers of individuals covered by the social security system must notify the agencies of health insurance, statutory pension

[2] For the time period covered applies the DEVO (Datenerfassungsverordnung) – Verordnung über die Erfassung von Daten für die Träger der Sozialversicherung und für die Bundesanstalt für Arbeit vom 24.11.1972, BGBl. I, 2159ff and the DÜVO (Datenübertragungsverordnung) – Verordnung über die Datenübermittlung auf maschinell verwertbaren Datenträgern im Bereich der Sozialversicherung und für der Bundesanstalt für Arbeit vom 18.12.1972, BGBl. I, 2482ff .

schemes and the Federal Employment Services (BA, "Bundesanstalt für Arbeit") about every new employment and every release. In addition, at the end of each year, i.e. on 31 December, an annual control notification is required for each employee that stays with the company. The information on when and how long people were employed and their salaries are collected in very fine detail and with high accuracy, since the insurance agencies need the information to calculate the insured employees' insurance entitlements. From these notifications, the Institute for Employment Research (IAB "Institut für Arbeitsmarkt- und Berufsforschung") of the BA compiles and continually updates a data file for each person covered (henceforth called "IAB data"). This file contains longitudinal information about all jobs of a person which were covered by the social security system of the FRG[3].

Table 1. Variables about employment episodes and transitions between employment episodes in IAB- and GHLS-data

IAB	GLHS
Social security number	Social security number *
Start date (Day, Month, Year)	Start date (Month, Year)
End date (Day, Month, Year)	End date (Month, Year)
Daily gross income on which social security contributions are being paid	Monthly income gross or net
Profession (three-digit code)	Profession (three-digit code + free text)
Professional position (code)	Professional position (code + free text)
Full or part time	Full or part time
	Exact working hours
	Fix-term contract?
	Required qualification
	Kind of contract
	How was employment obtained?
	Improvement respective to previous employment?
	Employment change voluntary or involuntary?
	Reason for employment change
	Number of employees in company
Company code	Company code
Branch	Branch (code + free text)
	Company name and place (free text)
	Company type

*Social security numbers were asked in the GLHS from those who agreed to have their records matched to the IAB-data

Survey reports are from the German Life History study's ("GLHS") subproject "Education and Careers of individuals born in 1964 and 1971 in West Germany", directed by Prof. K. U. Mayer at the Max Planck Institute for Human Development in cooperation with the Institute for Employment Research (IAB)[4]. The study collects data to describe and explain socio-economic processes and structures in the context of individual decisions and societal institutions and their

[3] About 80% of jobs in former West Germany are covered by the social security system (Bellmann et al., 2002) .
[4] Here, Hans Dietrich und Stefan Bender were responsible for the study.

changes over time. A representative sample of 2900 people born in 1964 and 1971 in West Germany – including German-speaking foreigners – was interviewed retrospectively in a CATI or CAPI[5] interview about their entire employment careers or employment episodes (Mayer and Brückner, 1989; Wagner, 1996).

The two data files contain different variables about each job episode and the transitions between them or into episodes without employment (see table 1). As identifying information, we will use the social security number and the start and end dates of each job episode. The social security number is used as a barring criteria that narrows down the search space: we compare only registered and reported episodes that have the same social security number and therefore concern the same person[6].

2.2 How Job Episodes and Their Start and End Dates Are Generated in the IAB File

In the IAB data file, every record represents a notification - either at the beginning of an employment, at the end of an employment, or the control notification on December 31. Notifications are related to the individual by their value on the social security number variable. Continuous employment episodes in a company can be created by using the unique company code which the Federal Employment Services assign to any company or employer. A change in company codes between successive records can be interpreted as an individual's transition into another job, and the continuity of company codes as a continuation of the same employment in the same company. We, therefore, merge successive records without time gap between them and identical company codes into an episode of continuous employment with one company - a "company episode".

How do the Federal Employment Services assign company codes to companies? A company is characterized as "smallest, legally independent, local unit of a work place as conceived by the Federal Statistical Office" (Brixy, 1999). This principle, however, cannot be applied universally. For example, branches of one company within the same community that belong to the same branch may apply for a common company code, provided that they carry out the notification from the same place (Bender and Hilzendegen, 1995). On the other hand, one *branch* can get more than one company code.

[5] Computer Assisted Telephone Interview und Computer Assisted Personal Interview.

[6] Matching individuals by the social security number is highly reliable: the digits are not random but contain meaningful information, and the last digits are produced by a checksum algorithm. If a respondent reports a wrong number or the interviewer records the number incorrectly, this will not lead to the wrong person being selected from the IAB file but to a complete match failure and the exclusion of the person from the match sample.

A new company code is generally given out whenever the owner of a company changes, unless the company is transferred to a family member, in which case a new company code only is given out if the health insurance agency demands it. Whenever a company changes its legal status, a new company code may be applied. If an owner closes down his company and establishes another one, he may use the old company number.

When changes in company codes do not correspond to a change for the employee that could be called a job or company change, respondents most likely won't report one. Three constellations might be most problematic (Fritsch, 1997) when interpreting company code changes between records longitudinally as changes in an individual's jobs:

- If a large company has more than one company code, a company code can change between notifications if the administrative task of making the notification is shifted to another place within the company.
- If owner or legal status have changed between notifications, a new company code may be used.
- If two or more branches of one company in the same community originally had the same company code, and one is being taken out of this (administrative) location – i.e. through outsourcing - , the company code for employees of this branch changes.

Fritsch (1997) suggests that this might be mainly a problem in the public sector, in large companies, in companies that belong to the social security branch and for non-profit organisations.

2.3 Construction of Employment Episodes in the GLHS

Starting with their first employment after completing primary education, respondents report their employment history retrospectively as a chronological sequence of employment episodes separated by times of unemployment or labor market inactivity, changes of company, professional position, income, activity or work time arrangements. We generate company episodes by merging employment episodes not separated by times of unemployment, times of labor market inactivity or company change.

How do people report their employment episodes? The employment episodes are based on the autobiographical recall of the respondents. They first have to describe their main activity during a time period as employment. Second, they have to reconstruct changes and describe them as end of employment, as a change between companies or a change within a company. Third, they have to date start and end of the episode thus reconstructed down to a month. In autobiographical memory, employment information is thought to be represented as sequence of employment episodes and interruptions that reflect the personally and socially meaningful and

consequential categories of the respondents' (work) biographies. These episodes do not necessarily correspond to what social scientists want and how they are represented in the IAB-data, which can lead to discrepancies. In addition, autobiographical memory often is incorrect, especially about calendar dates, which also leads to discrepancies between reported and registered episodes. Since autobiographical remembering tends to simplify and conventionalise the life course and adjust it to present individual and normative work-related roles and expectations (Barsalou, 1988; Conway and Pleydell-Pearce, 2000; Middendorf, 2000; Neisser and Fivush, 1994), a bias might be introduced here. Empirically, it has been shown (Paull, 2002; Reimer, 2003) that respondents retrospectively:

- omit and insert episodes of employment,
- temporally stretch, reduce and move employment episodes in both directions (towards earlier or later dates),
- omit transitions or merge employment episodes into a longer sequence of continuous employment if they adjoin directly or are separated only by a brief interval of unemployment/labor market inactivity, or by very brief other employment episodes,
- insert transitions or separate continuous company episodes into two or more episodes directly adjoining each other or separated by brief intervals of unemployment/labor market inactivity, or by very brief other employment episodes,
- describe a transition between employment episodes that occurred within a company as transition between companies or vice versa.

The selective omission of shorter events that do not fit into life's overall logic and the merging of similar episodes that are separated by unremarkable changes will be the most widespread memory errors. Describing transitions erroneously as within or between companies is especially likely when administrative structures are unclear or changing themselves; errors are equally likely in either direction (towards earlier and later dates) (Reimer, 2003).

Dating start and end of an episode down to a month relies on reconstructive inferences: people infer calendar dates of events from their connections with other biographical events and in reference to a few landmarks (such as birthdays or public holidays) of which exact dates are known. Dating errors are mostly small (a few months), or respondents err by exactly one, in some cases two or three years. Discrepancies usually are unsystematic, the erroneously reported date is as likely to be too early as too late. (Brown et al., 1985; Friedmann, 1993; Larsen et al., 1996).

Reconstruction errors do not occur equally often for all respondents; factors associated with persons or employment careers make the recall task easier for some cases, that therefore produce less discrepancies. Especially at risk are individuals with eventful or unconventional

careers that are harder to remember and are hard to be reported smoothly as a sequence of company episodes.

2.4 The Matching Procedure

A procedure for the matching of the retrospective employment history of the GLHS with the IAB register data has to take the following possible sources of discrepancies into account:

- respondents omit or insert entire episodes of employment,
- respondents misdate start and/or end of an episode in one or the other direction,
- respondents merge episodes separated by unremarkable transition, "ironing out" shorter periods of unemployment or labor market inactivity,
- respondents reconstruct company changes without any corresponding change in company codes,
- company codes change without any corresponding report of a company change.

We developed a five step matching procedure that allows - in addition to perfect matches - also *time-liberal* matches and *multiple matches* (see figure 1):

step 1: *Perfect Match*
Matches episodes with identical start and end dates

step 2: *Time-Liberal Match*
Matches episodes with start or end dates that differ up to 2 months

step 3: *Multiple Match IAB -> GLHS*
If an IAB-episode and a GLHS-episode start at the same date (plus/minus two months), and another IAB-episode and the same GHLS-episode end at the same date (plus/minus two months, the multiple shorter IAB-episodes are matched to the one GLHS-episode.

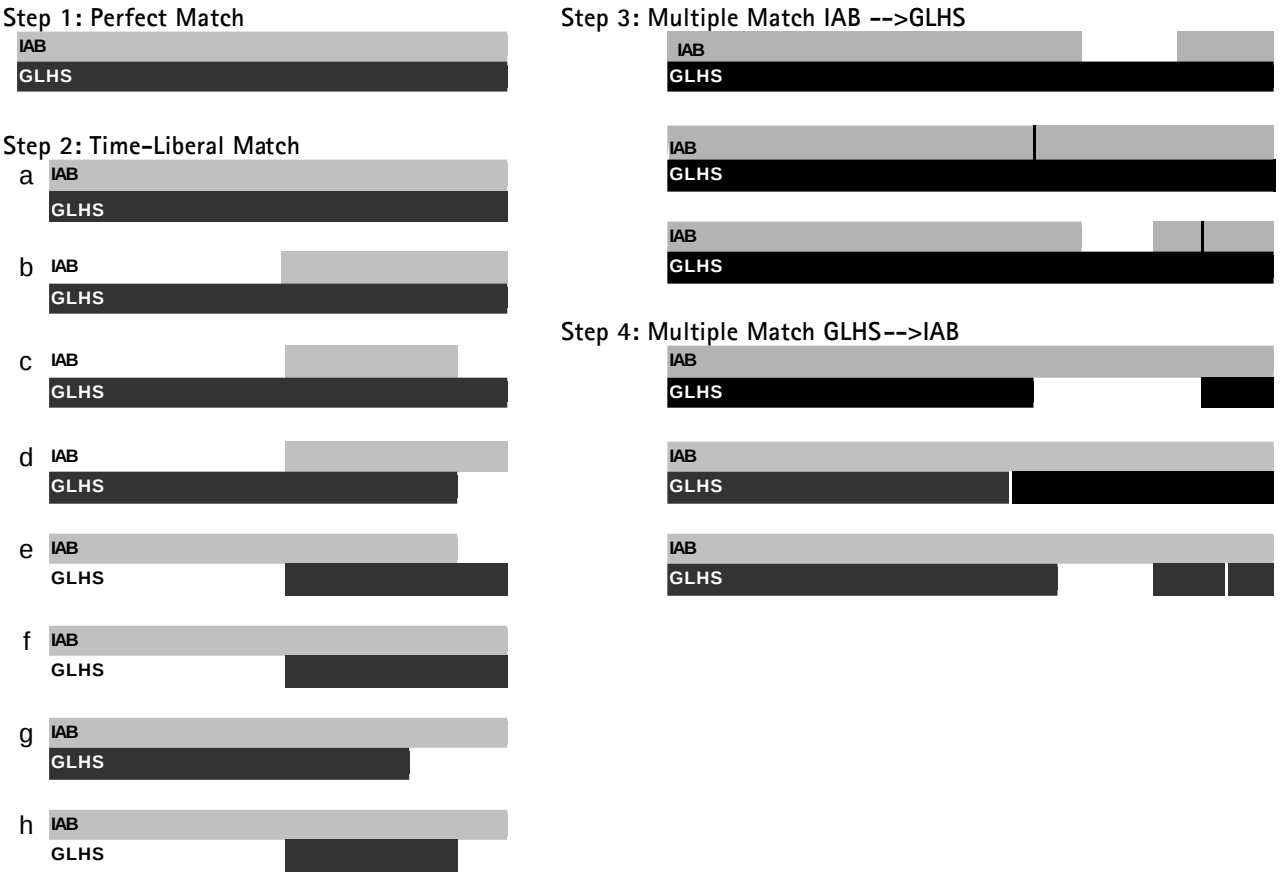
step 4: *Multiple Match GLHS -> IAB*
If a GLHS-episode and an IAB-episode start at the same date (plus/minus two months), and another GLHS-episode and the same IAB-episode end at the same date (plus/minus two months,) the multiple shorter IAB-episodes are matched to the one GLHS-episode.

step 5: *Match Failure*
Episodes unmatched by step 1 to step 4 are classified as non-matches.

Step 2 (Time-Liberal Matches) matches episodes that have different start and end dates due to small and unsystematic memory errors. Step 3 matches two or more registered episodes to one recalled episode, because respondents sometimes merge shorter episodes into one long episode by omitting transitions or "ironing out" shorter episodes of unemployment in between

employment episodes. This step also takes care of the cases where a registered company code change is not reported as company change by the respondent. Step 4 does the opposite and matches more than one recalled episode to one registered episode, since respondents occasionally insert company changes or unmerge episodes without a corresponding change in form code. In step 3 and 4, any number of episodes can be matched to one other episode, and gaps of any length between them may occur. Multiple matches run contrary to the Fellegi-Sunter-assumption that every record must be matched to one and only one other record.

Figure 1. Schematic representation of perfect, time-liberal and multiple matches



3 Matching Employment Episodes – Data Preparation and Match Rates

3.1 Sample

Of the original sample representative of people born in West Germany in 1964 and 1971 (n= 2909), we include employment episodes of the 636 individuals that can be matched accurately by their social security numbers[7]. This match sample is therefore a selection of those that agreed to the data linkage, have been covered by the social security system at least once and were able and willing to provide their social security number correctly. Table 2 gives an overview of some demographic characteristics of the original and the match sample. Differences in birth year and gender are minimal, but women of both cohorts are underrepresented and men of the 1964 cohort overrepresented. The lower educated (primary education) are underrepresented, those of intermediate and higher education are overrepresented.

Table 2. Differences between original and match sample

	Original sample		Match sample	
	n	%	n	%
Gender and Birth Cohort				
men	1530	52.6%	343	54.0%
women	1379	47.4%	292	46.0%
cohort 64	1474	50.7%	311	49.0%
cohort 71	1435	49.3%	324	51.1%
cohort 64 men	753	25.9%	173	37.2%
cohort 64 women	721	24.8%	138	21.7%
cohort 71 men	777	26.7%	170	26.8%
cohort 71 women	658	22.6%	154	24.3%
Highest Schooling				
No degree	53	1.8%	4	0.6%
Special education	15	0.5%	2	0.3%
einf. Hauptschule/ POS up to 8 years	509	17.5%	79	12.4%
qualif. Hauptschule	313	10.8%	50	7.9%
Realschule/Mittl. Reife/ POS up to 10 years	1045	35.9%	261	41.1%
Fachhochschulreife	96	3.3%	20	3.1%
Abitur/Fachabitur	743	25.5%	201	31.7%
Missing/Other	135	4.6%	18	2.8%
Total	2909		635	

[7] Bender et al. (2001) report on the linkage of the entire sample on the basis of individuals.

The differences may arise because the overrepresented groups have more often had jobs covered by the social security system. Also, the underrepresented groups might be less inclined to agree to the linkage or have greater difficulties in providing the correct social security number. The differences could bias our conclusions if these factors are related to outcome and success of the matching procedure. One might speculate that more conventional careers covered by the social security system can be matched more easily to social security register data, or that those individuals that are willing and able to report their social security number also reconstruct their careers more reliably. In this case, our observed match rates would be too optimistic. There are no empirical results, however, to confirm or disconfirm such speculations.

3.2 Employment Episodes in IAB and GLHS

A number of preparatory data transformations were necessary to make the episodes and identification information comparable. Since these constitute a part of the data generation process, table 3 reports in detail all necessary measures in both register and survey data. Single case studies were conducted to make sure these were appropriate. Table 4 gives an overview of the episodes in both data files after the transformation.

Table 3. Preparatory data transformation

	IAB	GLHS	Transformation
Employment episodes	Jobs covered by the social security systems (excludes civil servants (civil servants), self-employed, farmers; only jobs with a monthly income above 630,- DM; includes some traineeships)	Jobs for pay that constitute a person's primary activity	**IAB:** delete job training episodes **GLHS:** delete episodes with professional positions as self-employed, civil servants, helping relatives and farmers Flag episodes with a monthly income below 630 DM (n=4).
	Includes employment put on hold such as maternity leave, sick leave, or military services	Excludes times of maternity leave, sick leave, or military services	**IAB:** delete if income = 0
	Excludes employment with non-German employers		**GLHS:** delete if job outside Germany
Time period covered	1990-1997*	First job to time of interview (1998 or 1999)	**GLHS:** truncate at 1/ 90 and 12/ 97
Minimum duration of job episode	1 day	3 months	**IAB:** delete episodes shorter than 75 days **GLHS:** delete episodes shorter than 3 months Exception: episodes that start 1/90 or end 12/97 since these might be shorter due to the truncation of the period
Dates	Daily	Monthly	**IAB:** Transformation of daily to monthly dates according to the following rule: Start of episode: 1. - 15. - retain month 16. – last of a month: insert subsequent month End of episode: 1. - 15. of a month: insert previous month 16.- of a month: retain month Episodes of negative or zero duration: manual correction of dates
Minimum duration of gaps between episodes	1 day	1 month	**IAB:** temporal gaps between episodes within the same company are filled in if < 30 days of duration

*The IAB-data does cover the period up to the present; data for the time after 1997 was not available yet for external use at the time of data transfer.

Table 4. Episodes in IAB and GLHS

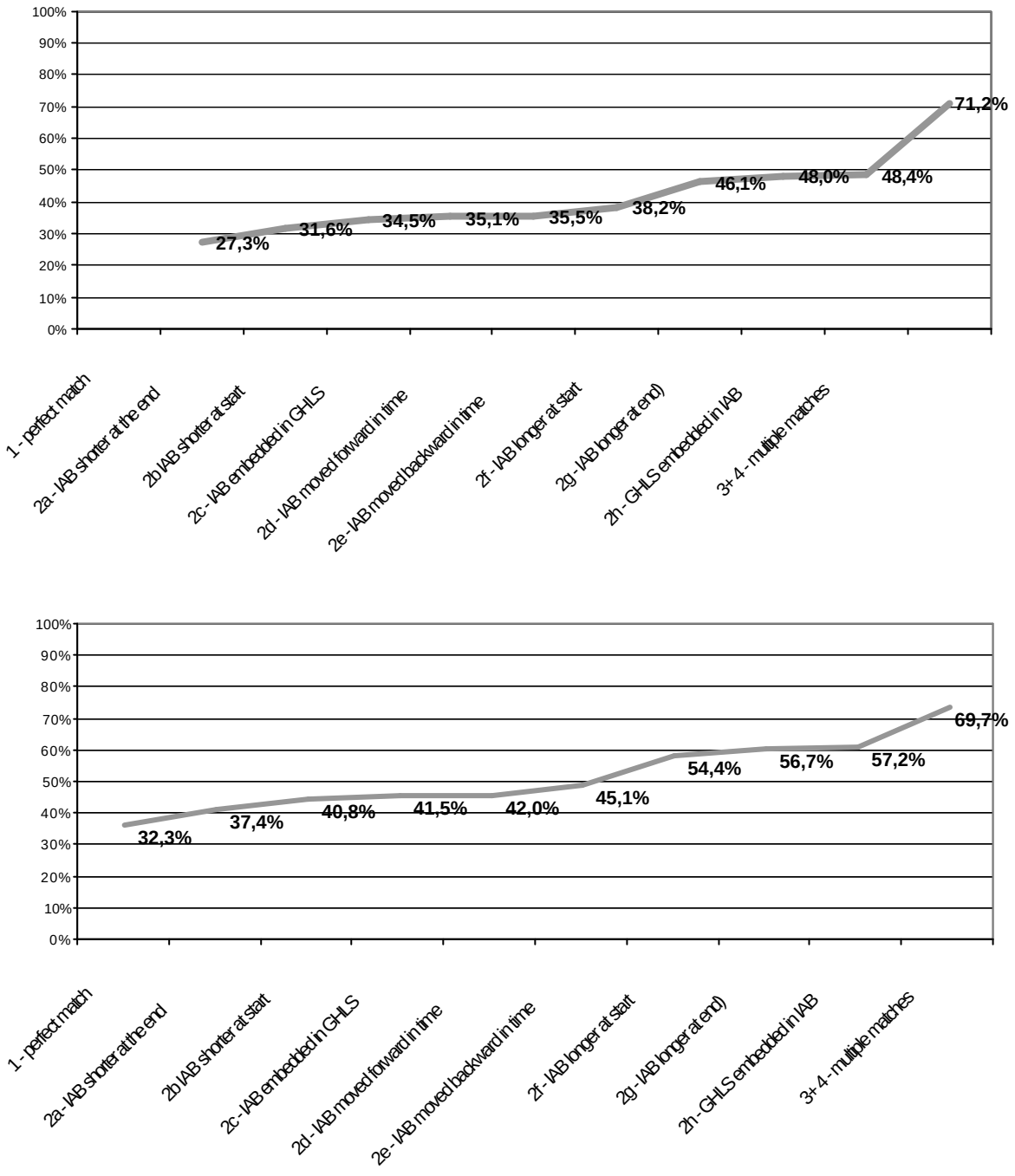
		IAB	GLHS
Episodes			
N Episodes		1255	1062
Duration*			
	Average (Months)	27.51	33.64
	Range (Months)	1 bis 96	1 bis 96
N Episodes or duration...*			
	1-3 Months	91 7.30%	69 6.50%
	3-12 Months	395 31.50%	264 24.90%
	13-24 Months	270 21.50%	196 18.50%
	> 24 Months	499 39.80%	533 50.20%
Persons			
N Episodes**		0-8 Episodes	0-7 Episodes
Time spent in employment**			
	Average (Months)	54.38	56.26
	0 Month	42 6.60%	56 8.81%
	1 Month – 3 Years	161 24.40%	13 20.90%
	3-5 Years	123 19.40%	113 17.80%
	5-7 Years	309 48.70%	333 52.20%
		635	635

* Excluded: persons without any employment
** Included: all 636 persons in the match sample

Table 4 shows that the episodes differ quite substantially between the two data files. The sequences reconstructed in the survey contain a much smaller number of employment episodes than the registered sequences, and these episodes are significantly longer on average. This is mostly due to a reduction in short and medium-length episodes of up to 2 years, and an increase in episodes of more than 2 years relative to the IAB-data file. This suggests that retrospective self-reports selectively underreport short episodes or merge them into longer episodes.

Similarly, the time spent in employment differs substantially. Careers with little time (up to 5 years, about 2/3 of the time period) spent in employment are underrepresented in the survey reports, while careers with more than 5 years are overrepresented. An exception are careers with no employment whatsoever; these are somewhat more frequent in the recalled sequences.

Figure 2. Match rates of the five-step match rule (% of matched episodes)



above: for IAB- episodes (registered episodes)
below: for GLHS-episodes (recalled episodes)

This can be seen as a result of the respondents' tendency to simplify and conventionalise their careers: individuals with few and short and therefore probably uncharacteristic and meaningless employment episodes eliminate these in their retrospective reconstructions, while persons who spend a lot of time in employment tend to prolong their job episodes by "ironing out" shorter

interruptions. Such aggregate level comparisons, however, give no information about individual correspondence between registered and reconstructed employment episodes.

3.3 Success of the Matching Procedure

We now apply our five-step match rule to the data using an SPSS-algorithm, and examine the resulting match rates, that is: the percentage of matched episodes from one data set. First, we look at the match rates at each step separately. The upper part of figure 2a shows the percentage of matched registered episodes from the IAB-data; the lower part shows the percentage of matched recalled episodes from the GLHS. These percentages are identical for the first two match steps (perfect and time liberal matches), only the multiple matching in steps 3 and 4 leads to different match rates for episodes from the two data sets.

With only perfect matches, the match rate for both registered and reported episodes is not even a third (step 1). Allowing in addition all forms of time-liberal matches (steps 2a to 2h), match rates go up to about half of all episodes. Multiple matches (steps 3 and 4) eventually lead to match rates of around 70 %. The matching of multiple (up to three) IAB-episodes to one GLHS-episode occurs much more frequently than the other way round. This is likely due to the tendencies towards retrospective simplification and conventionalisation that more often result in the merging of episodes and the ironing out of shorter interruptions than in the inserting of interruptions or transitions.

Next, we examine which episodes of which persons are matched more successfully. The left part of table 5 shows the match rates for men, women, the two birth cohorts and episodes of different length and for careers with different numbers of registered and reported episodes. Both registered and reported episodes of the younger birth cohort have a lower match rate, and episodes reported by women have lower match rates than those reported by men, but not episodes registered for women. The *number* of registered episodes in a sequence has no influence on the match rates, but there is a weak tendency for careers with more reported episodes to have lower match rates. The influence of episode *length* is more pronounced in both directions: shorter registered and reported episodes have lower match rates.

Table 5. Match rates for different kinds of episodes registered for and reported by different persons

	Episodes					Persons							
	IAB- episodes		GLHS-episodes			IAB-->GLHS				IAB-->GLHS			
						% of matched episodes in the sequence				% of matched episodes in the sequence			
	n	match rate	n	match rate		n	all	some	none	n	all	some	none
all	1255	71.6%	1062	69.7%		364	61.4%	18.0%	20.6%	355	61.3%	20.4%	18.3%
men	699	71.7%	570	71.7%		198	62.5%	18.9%	18.9%	198	64.9%	20.0%	15.1%
women	556	71.5%	492	68.6%		166	60.1%	17.4%	22.5%	157	57.3%	20.8%	21.9%
cohort 64	640	77.3%	519	75.0%		219	71.8%	11.8%	16.4%	219	72.3%	13.9%	13.9%
cohort 71	615	64.7%	543	64.8%		145	50.3%	24.7%	25.0%	145	49.8%	27.2%	23.0%
number of episodes													
1	226	72.1%	272	76.5%		163	72.1%		27.9%	208	76.5%	0.7%	22.8%
2	380	68.9%	386	65.0%		109	57.4%	23.2%	19.5%	99	51.3%	30.6%	18.1%
3	288	72.6%	210	72.9%		53	55.2%	32.3%	12.5%	33	47.1%	45.7%	7.1%
4	216	74.5%	124	68.5%		29	53.7%	37.0%	9.3%	11	35.5%	58.1%	6.5%
5	105	61.9%	45	71.1%		10	37.0%	44.4%	18.5%	4	30.8%	53.8%	15.4%
6-8	40	82.5%	25	56.0%									
episode duration					time spent in employment (years)								
1-3 months	48	52.7%	69	52.2%	up to 1	34	48.6%	7.1%	44.3%	38	71.7%	9.4%	18.9%
3 - 6 months	107	69.5%	118	68.6%	1 - 2	24	55.8%	14.0%	30.2%	22	55.0%	22.5%	22.5%
6 months to 1 year	166	68.9%	146	69.2%	2 - 3	26	54.2%	16.7%	29.2%	24	60.0%	12.5%	27.5%
2 years	190	70.4%	196	69.9%	3 - 4	27	50.0%	29.6%	20.4%	29	56.9%	23.5%	19.6%
3 years	118	76.1%	149	71.1%	4 - 5	40	58.8%	27.5%	14.5%	34	54.8%	25.8%	19.4%
4 years	83	66.4%	100	70.0%	5 - 6	45	54.2%	27.7%	18.1%	36	50.0%	33.3%	16.7%
5 years	59	79.7%	75	70.7%	6 - 7	50	60.2%	20.5%	19.3%	47	51.1%	26.1%	22.8%
6 - 7 years	122	84.1%	209	78.0%	7 - 8	118	82.8%	9.1%	8.4%	125	74.0%	13.6%	12.4%

The right part of table 5 looks at match success in a different way. For each person, we look at the percentage of all episodes that they reported or that were registered for them which could be matched: all, some or none. Clearly, match difficulties do not affect all person alike: both ways, the percentage of careers where *all* registered or reported episodes are matched is about 60 %. Of the remaining sequences, in about 20 % *some* episodes can be matched while others cannot, and in the remaining 20 % of careers, *no* episodes at all have found a match. Women's registered careers do contain slightly more often no matches and slightly less often are entirely matched. This tendency is more pronounced for women's reported careers. As for birth cohorts, both registered and reported careers of the younger birth cohort contain more often no matching episode and less often can be partially or entirely matched.

With an increasing number of registered and reported employment episodes, the percentage of entirely matched careers decreases markedly, with a corresponding increase in partly and unmatched careers. For the registered episodes, the more time respondents have spent in employment, the higher is the percentage of careers where all episodes are matched. The same tendency exists for the reported episodes, but less pronounced.

3.4 Qualitative Analysis of Match Failures

In order to better understand the reasons when and why our match rule failed, we carried out a qualitative analysis of 19 individual careers where no employment episode could be matched. We selected males and females from both cohorts that had different numbers of registered and reported episodes and displayed their careers graphically as lines that extend over a grid representing the months and years of the time period covered – analogous to the schematic careers depicted in figure 1. We looked at all information available in the GLHS and IAB data about the reported episodes that did not find a match and about the transitions between them. Since the survey asked also about people's activities during the times they were not employed, we also looked at what they report for the times when they were registered as employed but did not report this episode in a way that allowed to a match.

Five causes could be identified that repeatedly lay at the heart of the discrepancies; Appendix 1 shows graphical depictions of typical cases for each cause.

- Unclear transitions between training episodes and employment episodes

Respondents 1a, 1b and 1c report being in job training during times when they were registered as employed. Respondents 1b and 1c report a subsequent employment with the same company they were trained in. In the logic of the social security system, someone can legitimately be both in training and covered by the social security system as an employee. Such times may also seem ambiguous to a respondent in retrospect, but in the GLHS, these two states are mutually exclusive. For example, respondent 1b had to report his "Volontariat" (a practical training phase for journalists at a newspaper or news agency) either as a job or as training. Obviously, the training aspect seemed more prominent to him. Another possibility in such cases is that the training company did not report the status change when an apprentice or trainee was given a job; this became only mandatory in 1992 (Bender, 1997). This source of discrepancies could be responsible for the lower match rates of those born in 1971, since they are much more likely to have an episode of job training during the covered period.

- Inappropriate selection of episodes during the preparatory data transformation

As reported in table 4, we prepared the two data sets for the comparison by excluding some episodes from both files that by definition of the concepts and purposes of the respective data sets could not have a counterpart in the other file. For example, using the self-reported professional status, we deleted all the employment episodes during which respondents were civil servants, self-employed, farmers or helping a family member from the GLHS data set, since such jobs are not covered by the social security system and therefore cannot possibly show up in the IAB-data. Cases 2a and 2b show clearly that this also lead to the exclusion of episodes during which actually social security payments have been made - the self-reported professional status was incorrect. In case 2a for example, an episode as self-employed head waiter could have been matched in step 3 to the IAB episodes.

- Misdating in the GLHS data set

Respondents 3a and 3b report employment episodes that, from the codes for profession and branch, suggest that they correspond to registered episodes. But they date start and end of an employment episode by more than two months too early or too late respectively: 3a by three months, 3b by exactly one year. Time liberal matches only permit discrepancies up to 2 months and is therefore not liberal enough to match these episodes. A more liberal match rule, or one that also permits misdating by exactly one year, would have matched these episodes at step 2.

- Periods of continuous employment are segmented into discrepant company episodes

According to both data sets, respondent 4a reported employment for identical time periods. The codes for profession and branch also suggest that he reported the same activities as the register records. But these time periods are structured differently into company episodes by retrospective recall and company code change: he recalls only one company change (accompanied by a change in activity) while the register shows three. Moreover, the reported transitions occur at a very different point in time than does any of the three company code changes. Two explanations are possible: first, the respondent erroneously reconstructs transitions as within company changes that actually were between companies; during data preparation, such transitions were deleted and the two adjoining episodes merged. Second, company code changes do not bring with them marked transitions in job related roles, activities

and therefore, in the survey, are not (and, in many cases should not) be recalled as company changes. This can occur when company codes change due to administrative procedures that have no corresponding impact on the individual. This is possible especially since the respondent 4a is a civil servant.

- No plausible explanation

Some cases remain where, like for respondent 5, neither of the diagnoses 1 to 4 seem adequate. Such discrepancies may result from a combination of dating errors and different reconstruction of transitions, that are not to be disentangled anymore in a way that can be related to the two data generation processes.

4 Conclusion and Outlook

4.1 What Have We Learned About Match Rates and the Causes of Match Failure?

To highlight the specific challenges of matching and linking data on the level of events or episodes, we matched retrospective survey reports and register data about jobs and tried to derive explanations for match failures on the basis of knowledge about the data generation processes at all stages. Permitting time-liberal and multiple matches, about 70 % of episodes can be matched.

Match failures mainly occur at three points in the data generation process:

- when respondents reconstruct their episodes from autobiographical memory,
- when they recall the identificatory information about the reconstructed job episodes erroneously (here: starting and ending dates),
- when they reconstruct the information that is used to prepare the data and make it comparable with the register data.

We showed that in retrospective reports, short episodes are omitted or merged, when transitions are not very marked or short times of employment interruptions are ironed out, leading to simpler, more conventional and stable careers. Less frequently, transitions and episodes are inserted. Another source for match failures is misdating or company code changes due to administrative changes that do not correspond with relevant changes for the respondent.

Match failure does not concern every respondent to the same degree. Those respondents that have to report on fuzzy transitions between company based training and employment are more

at risk of match failures. The same can be observed for those whose career is characterized by irregular or rapidly changing employment patterns.

4.2 Theoretical and Practical Implications

a) Match rates are dependent on the criteria of the matching rule. This has already been observed by Miller and Groves (1985). As a consequence, they suggest that matching should generally be done with a wide range of criteria and combination of criteria and the resulting match rates reported in detail. In line with this, Luks and Brady (2003) systematically vary their temporal match criteria in order to adopt the one with the highest match rate as the most appropriate. In our study, for example, the time liberal matches could have been more generous and permitted discrepancies up to 3 months or more. Considering the common memory error of misdating an event by exactly one year, only those cases with a discrepancy of exactly 12 months could have been matched while maintaining the stricter criterion of two months. The choice of criteria, however, should be set in accordance with the aims of the linkage. We chose 2 months because we wanted a more conservative match rule that in doubtful cases would rather decide against a match, thus placing a relatively strong emphasis on avoiding type B errors. Moreover, a more liberal criterion makes it more difficult to decide about how to represent a career in the final linked data set (see point e). The variability of match rates should be kept in mind when evaluating the feasibility of data linkage or the quality and usefulness of the resulting linked data set.

b) Additionally, match rates depend on how the data sets are prepared for the matching. Survey and register data are gathered by different agents for different purposes which results in different data formats or different definitions for similar concepts. Most data linkage operations, therefore, have to carry out preparatory data transformation. Here, episodes had to be deleted or merged, variables recoded or text variables coded. Data preparation too could be done using a range of criteria to compare the influence on the match rates. In our case, we have tried to select the most appropriate way and achieve an optimal balance between Type-A and Type-B-errors by looking extensively at single cases. But as the analyses have shown, inappropriate selection could not be entirely avoided, since the selection criteria are error-prone in the same sense as the identificatory information.

c) The choice of good identificatory information and the choice of a useful match rule requires an in-depth knowledge about the processes that generate both the objects and the

identificatory information in both data sets. Indispensable here is the collaboration with persons involved in the collection and administration of both data sets. They can provide invaluable insights about the official rules and the informal pragmatics of the data generation procedures (Bick and Müller, 1980).

d) We abandoned Fellegi's and Sunter's classical assumption that every object is matched to one and only one other object by permitting multiple matches. This has markedly increased the match rates, especially the matching of more than one registered episode from the IAB to one recalled episode from the GLHS. It, however, requires careful decisions about how to represent this time period in the final data set and about which detail variables for each episode and transition to eventually use. This touches upon the wider question about which data set contains the more valid information about objects (episodes) and their start and end dates - or the question about which data set represents social reality more adequately.

Generally, register data are considered to be more valid – especially regarding calendar dates – since they are not flawed and biased by memory errors or interviewer influences. On this assumption, register data are used to validate survey reports (i.e. Auriat, 1992; Cash and Moss, 1972; Mathiowetz and Duncan, 1988; Means and Loftus, 1991). When linking data sets, they frequently serve as the reference data set from which the objects are taken, while adding the values of variables from the other data set to them.

Our single case studies have shown that match failure indeed can be plausibly explained by recall errors in many cases, which would make is plausible to give a priori preference to the registered episodes. However, social reality is always represented with error in any kind of data. There is also reason to believe that errors in survey and register data are correlated: certain persons with certain careers tend to be represented with a lower validity in both register and survey data. An example from our study are large companies in the public sector: Fritsch (1997) assumes that here, the interpretation of company code changes as actual company changes for individuals can be problematic; while Becker (2001) and de Graaf and Wegener (1989) have shown that civil servants have quite often difficulties to recall positions and jobs reliably.

Whether and when it makes sense to evaluate one data set against the other depends again on the purposes of the linkage. As an example, consider the cases in which a company code change due to administrative reasons does not correspond with an actual change for the respondent. Since the purpose of the GLHS is not to track administrative changes of companies but actual changes in peoples' lives, match failures cannot be interpreted as errors or flaws in the survey data. Vice versa, it is not the purpose of the register data to track individuals lives longitudinally.

This concerns rather conceptual correspondence than differential validity. Miller and Groves (1985) talk about register and survey data as alternative ways to represent the same aspects of social reality.

Practically, this would imply the creation of an integrated "reconciled" data set instead of prioritizing one over the other on the basis of an assumed greater validity. Paull (2002) created such an integrated reconciled data set in a study where she matched reports about job episodes from two overlapping panel waves in the British Household Panel Survey (BHPS). She compared the reconciled data with two others, that gave priority to one of the panel waves, and a third, that excluded all doubtful cases from the final data set. The main advantages of the integrated data set were that more cases could be included and memory errors were counteracted. However, a reconciled data set probably will require manual matching in addition to the automatized match rule.

e) Matching decisions can be made "manually" by individuals that look at single cases and careers, or by automatic rules and algorithms. While manual matching is only feasible for a small number of cases, automatic rules can use only a small number of variables as identificatory information and are notoriously weak in "world knowledge". They will inevitably fail to do justice to some cases and intricacies of the differences in data generation processes which leads to errors of Type A and Type B both at the preparatory stage and during the matching itself. Of course, also individuals will commit errors and introduce biases. To our knowledge, the only study that compares match rates of automatic procedures and individuals is the one by Miller and Groves (1985). The authors conclude that match rates from individuals resembled more closely the computerized matching rules with relatively liberal criteria.

In this methodological study, we had time and resources to check and revise the automatized preparatory steps as well as the matching steps again and again by looking at single cases. Also, at some points, inappropriate consequences of the automatized matching rules for some special careers were corrected manually. We were therefore able to combine both the advantages of automatic rules and individual matching. We remain sceptical that the human being will ever be completely dispensable with when matching data, at least when information as complex and idiosyncratic as ours is concerned. As we have seen, the majority of careers that contain few episodes, so that for most persons, the matching was unproblematic. We assume that optimal matching results will require both: on the one hand, an automatized rule to deal with the bulk of relatively simple and straightforward cases, and on the other hand, individuals that

continuously check the results for appropriateness and deal with the more complex and challenging cases.

5 Literature

Auriat, N. (1992). Who forgets? An analysis of memory effects in a retrospective survey of migration history. *European Journal of Population 7*, 311-342.

Barsalou, L. W. (1988). The content and organization of autobiographical memories. In: E. Winograd (ed.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory*. New York: Cambridge University Press.

Becker, R. (2001). Reliabilität von retrospektiven Berufsverlaufsdaten - ein Vergleich zwi-schen der Privatwirtschaft und dem öffentlichen Dienst anhand von Paneldaten. *ZUMA-Nachrichten* 49, 29-56.

Bellmann, L., Bender, S. & Kölling A. (2002). Der Linked Employer-Eployee-Datensatz aus IAB-Betriebspanel und Beschäftigtenstatistik der Bundesanstalt für Arbeit. In: G. Kleinhenz (ed.), *IAB-Kompendium Arbeitsmarkt- und Berufsforschung*. Stuttgart: Kohlhammer.

Bender, S., R. Brand & J. Bacher. (2001). Re-identifying register data by survey data: An empirical study. *Statistical Journal of the United Nations* 18, 373-381.

Bender, S. & J. Hilzendegen. (1995). *Die IAB-Beschäftigtenstichprobe als scientific use file*. Stuttgart: Kohlhammer.

Bender, S. (1997). Die IAB-Beschäftigtenstichprobe 1975-1990: Analysemöglichkeiten der anonymisierten Stichprobe. In: R. Hujer et al. (eds.), *Wirtschafts- und sozialwissen-schaftliche Panel-Studien - Datenstrukturen und Analyseverfahren - Sonderhefte zum Allgemeinen Statistischen Archiv*. Göttingen: Vandenhoeck und Ruprecht.

Bick, W. & P. J. Müller. (1980). The nature of process-produced data - towards a social-scientific source criticism. In: E.K. Scheuch (ed.), *Historical social research: The use of historical and process-produced data*. Stuttgart: Klett-Cotta.

Brixy, U. (1999). Die Betriebsdatei der Beschäftigtenstatistik der Bundesanstalt für Arbeit. *Beiträge zur Arbeitsmarkt- und Berufsforschung* 230

Brown, N. R., L. J. Rips & S. L. Shevell (1985). The subjective dates of natural events in very long-term memory. *Cognitive Psychology* 17, 139-177.

Cash, W. S. & A.l J. Moss (1972). *Optimum recall period for reporting persons injured in motor vehicle accidents*. Washington D.C.: National Center for Health Statistics.

Conway, M. A. & C. W. Pleydell-Pearce (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review* 107, 261-288.

de Graaf, Nan Dirk & B. Wegener (1989). (Un)reliability of job career and social resource data. *Unpublished manuscript.*

Fellegi, I. P. & A. B. Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association* 64, 1183-1210.

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin* 113, 44 – 66.

Fritsch, M. (1997). Die Betriebsdatei der Beschäftigtenstatistik - Ansatz und Analyse-möglichkeiten. In: G. Wagner (ed.), *Wirtschafts- und sozialwissenschaftliche Panel-Studien - Datenstrukturen und Analyseverfahren*. Göttingen: Vandenhoeck.

Larsen, S. F., C. P. Thompson & T. Hansen (1996). Time in autobiographical memory. In: D.C. Rubin (ed.), *Autobiographical memory.* Cambridge: Cambridge University Press.

Luks, S. & H. E. Brady (2003). Defining Welfare Spells: Coping with problems of survey responses and administrative data. *Evaluation Review* 27, 395-420.

Mathiowetz, N. A. & G. J. Duncan (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. *Journal of Business and Economic Statistics* 6.

Mayer, K. U. & E. Brückner (eds.) (1989). *Lebensverläufe und Wohlfahrtsentwicklung - Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929-31, 1939-41, 1949-51, Teil 1-3*. Berlin: Max-Planck-Institut für Bildungsforschung.

Means, B. & E. F. Loftus (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology* 5, 297-318.

Middendorf, E. (2000). Panta rhei oder der mentale Fluß von Tatsachen: Zur Reliabilität retrospektiv erhobener biographischer Ereignisse. *ZA-Nachrichten* 46, 58-71.

Miller, P. V. & R. M. Groves. (1985). Matching survey responses to official records: An exploration of validity in victimization reporting. *Public Opinion Quarterly* 49, 366-380.

Neisser, U. & R. Fivush (eds.) (1994). *The remembering self. Construction and accuracy in the self-narrative*. Cambridge: Cambridge University Press.

Paull, G. (2002). Biases in the reporting of labour market dynamics. Working paper of the *Centre for Economic Performance*. London: The Institute for Fiscal Studies.

Reimer, M. (2003). Autobiografisches Erinnern und retrospektive Längsschnitt-datenerhebung: Was wissen wir und was würden wir gerne wissen? *BIOS - Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen* 1, 27-45.

Wagner, M. (1996). Lebensverläufe und gesellschaftlicher Wandel: Die westdeutschen Teilstudien. *ZA-Information* 38, 20-27.

6 Appendix

Appendix 1. Illustrative graphical depictions of the five kinds of discrepancies