Schmollers Jahrbuch 130 (2010), 485 – 512
Duncker & Humblot, Berlin

The Impact of Cleansing Procedures and Coding Decisions for Overlaps on Estimation Results – Evidence from German Administrative Data

By Patrycja Scioch*

Abstract

Process-generated and administrative datasets have become increasingly important for labor market research over the past ten years. Their major advantages are large sample sizes and the absence of retrospective gaps and unit non-response. Nevertheless, the quality and validity of these types of data remains unclear, and a great deal of preparation and data cleansing is necessary before the data can be analyzed. Unfortunately, few researchers explicitly describe the cleansing procedures or coding decisions used for this purpose, thus leaving their impact on the results unclear. The present paper focuses on the variation in research results resulting from different cleansing and coding procedures. The paper uses the framework of data preparation proposed by Wunsch / Lechner (2008) as a benchmark, and induces variation by developing different cleansing procedures and coding decisions for overlapping and parallel observations. The descriptive results show that the data sets (resulting from the different procedures) show varying ranges of difference for some attributes related to time and personal characteristics. Similar results emerge from the subsequent analysis of treatment effects, which do not vary in overall shape but in magnitude, especially during the lock-in effect. In sum, the results indicate that the empirical findings of evaluation studies based on matching algorithms are fairly robust to variations in the underlying method of data preparation.

Zusammenfassung

In der Arbeitsmarktforschung ist in den letzten zehn Jahren die Bedeutung von administrativen Daten zunehmend gestiegen. Bedeutende Vorteile gegenüber Befragungen sind große Stichproben und das Umgehen von typischen Befragungsproblemen wie Erinnerungslücken und Antwortausfällen. Allerdings bleibt die Validität von administrativen Informationen unklar und es ist viel Datenaufbereitung und -bereinigung notwendig um die Daten für Analysen nutzbar zu machen. Leider geben nur wenige Forscher Einblick in ihre Bereinigungsprozeduren, womit deren Einfluss auf die eigentlichen Ana-

* I would like to thank Conny Wunsch and Michael Lechner for supporting the work related to this paper and for giving access to their program codes, as well as Dirk Oberschachtsiek and two anonymous referees for their comments. This paper is an extended version of Oberschachtsiek / Scioch (2009).

lyseergebnisse unbekannt bleibt. Diese Studie greift dieses Thema auf und untersucht die Variation von Forschungsergebnissen aufgrund von alternativen Bereinigungsprozeduren. Insbesondere werden die von Wunsch / Lechner (2008) angewendeten Bereinigungen für überlappende und parallele Beobachtungen in Individualdaten als Referenz genutzt und davon ausgehend weitere Bereinigungsprozeduren entwickelt. Diese verschiedenen Prozeduren führen zu Unterschieden hinsichtlich zeitlicher und individueller Charakteristika. Ähnliche Ergebnisse weisen die Analysen zu Maßnahmeneffekten auf, welche sich nicht im generellen Verlauf, sondern in der Höhe der Effekte – vor allem im Lock-in-Effekt-unterscheiden. Alles in allem weisen die Ergebnisse der Untersuchung darauf hin, dass empirische Resultate dieser Evaluationsmethode relativ robust gegen Änderungen der zugrunde liegenden Bereinigungsprozeduren sind.

JEL-Classification: C81, J68

Received: September 30, 2010
Accepted: January 21, 2011

1. Introduction

Process-generated and administrative datasets have become increasingly important in labor market research in Europe over the past ten years. While other countries like the USA used administrative data earlier in the evaluation of training programs (Ashenfelter, 1978; Angrist 1998; Mueser et al., 2007) or attempted to develop them for use in official statistics (Jabine / Scheuren, 1985) the development in Europe has been rather slow. Kluve et al. (2006), for example, report that in the late 1990s, most countries used survey data for labor market policy evaluation. He adds that over the past decade, this has changed and that now the vast majority of microeconomic evaluation studies in Europe (almost 75%) are based on administrative data. Particularly Scandinavian labor market research shows that register data can be a valuable source for empirical research (for example, Eliason / Storrie, 2006; Carling / Richardson, 2004; Røed / Raaum, 2003; Geerdsen / Holm, 2004; Hämäläinen / Ollikanen, 2004).

In Germany the number of studies has increased since 2000 thanks to the efforts of research groups from various institutes to utilize administrative data from the Federal Employment Office (Klose / Bender, 2000; Hujer et al., 2004; Lechner et al., 2004; Fitzenberger / Speckesser, 2007). Another cornerstone was the official report of the Commission on Improving the Informational Infrastructure between Science and Statistics (in German: Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik; KVI, 2001), which recommended the creation of Research Data Centers (RDC) and Data Service Centers (DSC). The first of these were established in 2004 to provide access to administrative data for research (e.g., the RDC of the Federal Employment Agency in the Institute for Employment Research). This new service resulted in a growing body of research based on

this type of data (for example, Lechner / Miquel, 2009; Bauer et al., 2007; Rinne et al., 2008; Fitzenberger et al., 2009).

In comparison with traditional survey data, register data cover a much greater number of observations. Administrative data[1] are therefore often used to overcome weaknesses of survey data such as attrition bias, reporting or collection bias, a lack of relevant comparison groups, and small sample size. The most important advantage of administrative data, however, lies in the possibility to merge data from different sources and different points in time.

Nevertheless, few studies to date have focused on the quality and suitability of administrative data for empirical research. For over 20 years, research has sought to assess and improve the quality and representativeness of survey data (e.g., Groves, 2004; Groves et al., 2004). This assessment constitutes one of the primary uses of administrative data (see e.g., Pyy-Martikainen / Rendtel, 2008; Reimer / Künster, 2004; Jenkins et al., 2005; van den Berg et al., 2004). Administrative data are not, however, generally used as a sole data source for research. Administrative data are also used to identify any differences between the use of register and survey data and to determine if one or the other source is superior (see, e.g., Blank et al., 2009; Rendtel et al., 2004; Hotz / Scholz, 2001).

With regard to process-generated data, the literature on the measurement, improvement, and validation of data quality is scant in almost all countries. Jabine / Scheuren (1985) defined goals for the use of administrative records in official statistics in the US, and Wallgren / Wallgren (2007) described the use of administrative records for statistical purposes in general. Aside from discussions of quality issues in producing statistics the issue of data quality has seldom been addressed in empirical evaluation studies, either on its own or as a part of broader evaluations. Johansson / Skedinger (2005) assessed data from the Swedish Public Employment Service and found systematic misreporting due to incentives to misreport disability status. Rendtel et al. (2004) analyzed the reliability of Finnish income data by comparing them with survey data and attempted to identify an appropriate measure of quality. For Germany, Fitzenberger et al. (2006) developed imputation rules to improve the education variable in a widely used administrative data set, while other studies on German data have focused on issues arising in the complex data generation process (e.g., Kruppe / Oertel, 2003; Engelhardt et al., 2008). Further studies show that administrative data are faced with similar problems to survey data, including missing values, overlaps, and inconsistencies. Jaenichen et al. (2005) and Bernhard et al. (2006) identify distinctive types of implausible cases in a German data set and discuss simple heuristics to handle these types of inconsistencies. Both of these studies focus on overlaps and gaps, and emphasize

[1] In the following register and administrative data are used synonymously.

the need for data preparation and data cleansing. More recent papers such as Kruppe et al. (2008), Fitzenberger / Wilke (2009), and Waller (2008) focus on the link between research results and data processing procedures. The first two studies deal with the different definitions of unemployment and possible effects on evaluation results, while the latter develops different correction procedures for the end dates of program participation and discusses the influence on estimation results, finding only small differences in the treatment effects caused by measurement error.

Due to the possibility of merging data from different sources and the fact that administrative data are not collected by the researcher directly, almost every study that uses administrative data must cleanse the data before conducting the analysis of interest. This issue becomes even more pronounced when parallel observations with (potentially) contradictory information are involved and when information on data quality is absent. In such cases, the cleansing and preparation of the data is indispensible to the empirical investigation. The modus operandi usually involves making specific decisions on cleansing procedures such as the use of certain rules or predominance criteria. However, data processing has seldom been subjected to systematic empirical investigation.

This study seeks to overcome this gap. By using an evaluation study as example and benchmark, this paper investigates the impact on evaluation results of different cleansing procedures[2] for overlapping observations in a merged administrative dataset. In this context, the robustness of the results is important since evaluation studies are usually related to policy interventions. The data I focus on here is of substantial interest and widely used in labor market policy evaluation in Germany. Previous studies in this field using the same data (e.g., Stephan, 2008) showed that the meaning and size of estimated treatment effects depend heavily on the choice of treatment and comparison group. This point (rules for selection of treatment and control groups) is therefore held constant, and the investigation focuses on the cleansing of record overlaps and inconsistencies between the different sources of this database. In a first step, I describe the data cleansing and coding approach suggested by Wunsch / Lechner (2008) and conduct a broad analysis of the training programs in Western Germany. In a second step, I propose variations in the cleansing procedure and analyze the effects of the variations on the point estimates within the evaluation framework by comparing the results obtained by each cleansing method with the results of the reference method. As reported by Waller (2008) in a similar study, I find no major differences between the effects; the main differences occur in the short run in the so-called lock-in effects. The results therefore emphasize that the empirical findings are robust to variations in the underlying cleansing procedure.

[2] In the following cleansing procedures and coding decisions are used synonymously.

The discussion of the analysis is organized in six subsections, which are structured as follows: In the next section, the database is described and problems are discussed that may occur when using this large administrative database with its wealth of information and large sample size on the one hand, and its inconsistencies and overlapping records on the other. Section 3 outlines the general framework and describes in detail the coding procedure proposed by Wunsch / Lechner 2008. I use this cleansing procedure as benchmark in all later sections to identify potential differences and discuss the later results. Section 4 describes the development of the new procedures, and Section 5 presents and discusses the descriptive statistics and point estimation results. Section 6 summarizes the main findings and concludes.

2. Database

The database used in this study is the Integrated Employment Biographies[3] (IEB) of the Institute for Employment Research (IAB) in Germany, which is a longitudinal data set merged from four distinct process generated data sources. The data cover nearly 80% of the total labor force in Germany and almost 100% of the employees eligible for social security benefits. Not included are civil servants, periods of self-employment, and periods of childcare leave. The sources of the data set are four administrative processes, which are linked by a unique identifier. Each of these sources offers a broad set of attributes and covers different periods of observation.

– The first data source is the *Employment Histories,* containing employment periods captured by the social insurance register going back to 1990 (marginal employment since 1999). Besides the start and end dates of employment, it also includes the employment status, personal characteristics such as gender, education, experience, age and nationality, and information about the job such as daily wage, occupational status, type of profession, region, and industry. Moreover it allows the merging of further information about the employer by using an establishment identifier and it allows adding information to individuals by using an individual identifier. Changes in territorial allocation are updated in current observations as well as in previous ones.

– The second data source contains data on spells of unemployment from the *Benefit Recipient History.* It has information on a daily basis on the amount and duration of the receipt of unemployment benefits, unemployment assistance, and subsistence allowances since 1990. Additionally, the source includes personal characteristics and statements on sanctions due to lack of

[3] A detailed description of the data structure can be found in Köhler / Thomsen (2009). For a brief description of the IEBS (which is a weakly anonymized 2% sample of the IEB), see Jacobebbinghaus / Seth (2007).

cooperation with the Public Employment Service (PES) or non-appearance at interviews with the PES staff.

– Most of the individual characteristics in the IEB data arise from the *Applicant Pool Data*, which contains information on job search spells since 1999. Apart from current marital state, nationality, health, education, and regional characteristics, the data set also comprises information about the last job, the desired job, and profession.

– Finally the data set on *Active Labor Market Program Participation* provides information on periods spent in subsidized schemes (e.g., training programs). Since 2000, all participation in employment, training, or job-creation measures has been recorded with start and end date, personal characteristics of the participant, and information about the program such as topics and success.

It is important to note that the sources are not cross-validated, which may create parallel observations (overlaps). It is possible for individuals to have several jobs at the same time or to be employed and searching for a new job simultaneously, or to be receiving benefits while looking for a new job or participating in labor market programs. These spells may be completely parallel, one may be embedded within the other, or they may overlap. The existence of parallel observations is twofold: it may offer additional information such as periods of subsidized employment, or it may cause problems when information is contradictory and the two observations cannot occur simultaneously. This may be the case, for example, if an individual is participating in a full-time training program and has a full-time employment observation parallel to this. In such cases, one must decide which data source to believe and choose–that is, which will be the subject of data processing and coding.

To combine the abundance of information into one manageable data set, a variety of characteristics have to be selected from each source and linked. Köhler / Thomsen (2009) elaborately describe the data integration and consolidation process, while Seysen (2009) describes the effects of changes in the mode of data collection on data quality. The IEB data are organized on a daily basis and therefore allow controlling for time varying covariates. Due to the huge size of the IEB, a 2.2% random sample[4] of the original IEB is used in this study, enriched with additional information from the four sources and a wide range of regional statistics added from INKAR[5] such as local unemployment rate, the share of foreigners, labor force participation rate, household income, and the share of long-term unemployed. As described above, the data

[4] The sampling procedure applies the same rules as used for the construction of the IEBS (see also footnote 3); for detailed information, please contact the author.

[5] Dataset of the Federal Institute for Research on Building, Urban Affairs, and Spatial Development.

are prone to parallel and overlapping observations. This is reflected in the increasing number of overlaps in the data over time. The 2.2% sample of the IEB has 34% overlapping observations in the period from 1990 – 2000. Since then, this number has increased to 49%, which means that decisions have to be made as to which observation to choose in almost half of the cases.

3. Benchmark

This study is rooted in what has become a broad body of research on the outcomes of labor market programs assessed at the micro level (e.g., Lechner et al., 2004; Biewen et al., 2007; Mueser et al., 2007; Osikominu, A., 2008; Fitzenberger et al., 2009). However, the focus here is not on the overarching framework of evaluation methods but on the use of appropriate methodologies to study the effects of data cleansing procedures for overlapping observations. In particular, I will refer to the study by Wunsch / Lechner (2008) as a benchmark and allow for variance in the methods of data cleansing. The effect of this variance on the outcome measures will be investigated.

This choice was made for four reasons. First, Wunsch / Lechner (2008) use a data base that contains administrative data and is already widely used. Second, the data is very complex in terms of its sources and generation, so advice on data cleansing and the question of the robustness of results may be useful. Third, the authors provided access to the majority of their program codes, which makes it possible to reconstruct the basic procedure in cleansing the data. Finally, this approach to data cleansing appears innovative and may therefore be of interest to other researchers and users of administrative data – in particular when using data with a high number of overlaps.

The idea of the investigation in this paper is to adapt a reference procedure including data cleansing and estimation of the interested outcome, and then to create variations in the underlying data cleansing procedure while holding the data set and the method constant as far as possible. This would allow identification of any difference in the outcome measure as a result of the cleansing procedure. With respect to the data structure, variation mainly occurs in the handling of parallel observations. In order to cope with this issue, the observations on the individual level are regarded within time spans for which the cleansing rules can be applied. Furthermore, since every observation has specific qualities, such as length and source of information, this information is used as the major characteristics of the data cleansing procedure. The cleansing aims to identify one valid state for each timeframe, and finally to transform the data into a panel data set.

Defining the data cleansing procedure

The data cleansing procedure consists of two parts. The first part concentrates on separating the longitudinal data into timeframes of two weeks. Each timeframe may then consist of several parallel or overlapping episodes of observations, which may differ in length and source. In order to isolate one state out of the parallel ones, sorting rules are applied to create an order of precedence in which, in the second part, one state can be selected. Two ordered sorting rules apply in the first part:

1. *Sorting Rule 1 (Length Priority):* First, all parallel episodes are sorted by length.
2. *Sorting Rule 2 (Source Priority):* If two or more parallel episodes have the same length (within the two-week timeframe), the respective data source is used as a proxy of the validity to order the observations.

Sorting the overlapping episodes by length and priority, and therefore into a particular order is the key to the entire cleansing approach. However, this investigation only concentrates on the second rule. This in turn means that Rule 1 will not be changed and that variance is only caused by changing the order of importance of the sources. To some extent, this affects the validity and reliability of the sources. In Part 2, after having ordered the episodes, only one general rule exists to select the final state:

Selecting Rule (Source Priority): Out of States 1 and 2, the final one is selected by applying predefined rules based on the priority of the respective source.

For the benchmark, the classification of priorities for sorting the sources follows the approach of the reference study and is referred to for the rest of this paper as procedure V0. In this procedure, participation in a labor market program is given the highest priority because it is at the heart of the evaluation design. Sources associated with payments (which are the Benefit Recipient History and the Employment History) are regarded as relatively reliable and follow in second and third place. The job search register, with a great deal of optional information, is considered to be less reliable with respect to the start and end dates and therefore has the lowest priority.

Figure 1 illustrates and describes this procedure. Imagine that the upper panel of the figure is an abstraction of an individual employment history that can be observed in the IEB. Each line represents an observation of a certain employment state (wage work, receiving unemployment benefits, employment search, etc.) with the start and end dates in parentheses. What one can see is that several parallel observations exist (some may be legally allowed, others not) coming from the same source of information, from different sources, or even from combinations of these. For example, it is legally possible to be employed and at the same time officially looking for a job (two sources, legal combination) but it is not permitted to be employed full-time and to receive

unemployment benefits at the same time. The legitimacy of some combinations may vary due to changes in the laws, which means that also the time has to be taken into consideration.

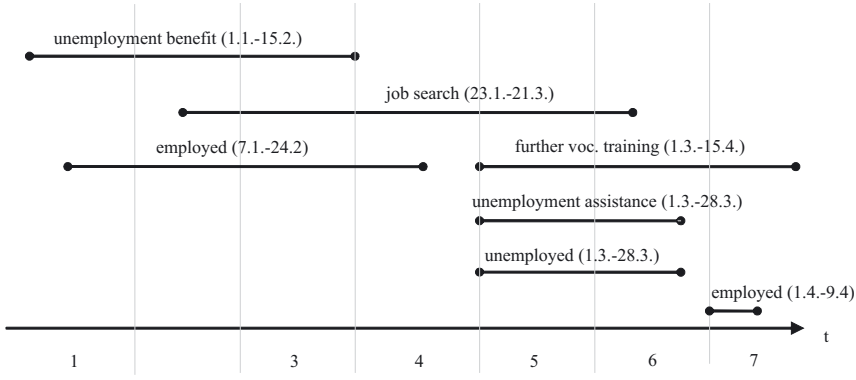
The *x*-axis represents time and is divided into (seven[6]) timeframes of two weeks each. These timeframes form the basis for the number of observations that will be isolated. Furthermore, seven observations are reported during the whole period of observation. As one can see, there is a benefit observation lasting from timeframe 1 until the end of timeframe 3 and an employment[7] observation beginning in timeframe 1 and ending in the middle of timeframe 4. The data cleansing procedure now aims to define one single unique employment status for each timeframe. This is displayed in the lower panel of Figure 1, which consists of a table that shows the transformation of the observations into the different states. Each column represents one timeframe. The rows contain the different states in that timeframe (e.g., timeframe 1 – see T1 – covers two states and timeframe 5 contains four states – see T5).

As mentioned above, the most important step in the data cleansing approach is sorting. Therefore, the order of states across the columns in Figure 1 is crucial. For example, the first column displays two episodes from different sources, one from the receipt of benefit source, and the other from the employment history. The first row contains the longest episode in the timeframe. If multiple episodes of the same length are observed, the episodes are sorted by heuristic routines (see timeframe 5). All episodes of participation in a training scheme are classified as having the highest priority, since their evaluation is the main point of interest in a program evaluation study. Episodes from the job search register are not associated with any type of payment (two possible states: searching and unemployed) and are therefore considered less valid and classified as lower priority.

After having identified the first two states in step one, these states are now processed in the second step of data cleansing. The final state is now selected by applying the selection rule from the first two episodes. Although for simplicity I discuss and display the priority of just four sources, the selection rule implies a significantly larger number of rules that define which state to choose (over 70 rules). Every source is divided into several states and every combination of overlaps is possible in the data: for example, participation in a degree course (program) is higher in priority than receiving a subsistence allowance when they are parallel, but not being unemployed and searching for a job (job search register) while receiving a subsistence allowance leads to the choice of subsistence allowance as the final state.

6 The number is for purely illustrative purposes.

7 For the sake of simplicity, the sorting of observations is described here in a highly aggregated manner (by employment status, etc.) irrespective of the particular values, which are used in the procedures.



	T1	T2	T3	T4	T5	T6	T7
State 1	benefit	benefit	benefit	search	training	training	training
State 2	employed	employed	employed	employed	assistance	assistance	employed
State 3		search	search		search	unemployed	
State 4					unemployed	search	
…							
Final state	benefit	benefit	benefit	employed	training	training	training

Figure 1: Example of an individual's history with overlaps

To demonstrate the choice of the final state, the example continues in the last row of the table in Figure 1. In timeframe 2 (column T2) episodes of observations from the unemployment benefit register and the employment history occur. Following the rules of priority, the first state is defined as the final state. Likewise, in period 5 (column T5) the final state (further vocational training) arises because unemployment assistance has a lower priority than participation in a labor market program. Note that the first state is not always chosen as the final state (see T4). If the source of state 2 has a higher priority than the source of state 1, the final state would be the one of state 2. This is displayed in timeframe 4, where the final state is employment, because being employed has a higher priority than searching for a job. As mentioned above, this is just a very simple description of the rules applied to illustrate the approach.

4. The Development of the Cleaning Procedures

To examine whether cleansing procedures have a noteworthy impact on estimation results, in the following, the benchmark procedure (V0) is modified in

different ways to develop new procedures. Subsequently the entire data cleansing and preparation process is done with the new procedures. This results in new evaluation samples that are compared with the benchmark sample of V0.

As described in Section 3, the procedure consists of two main sorting rules. Rule 1 orders the observations per timeframe by length, which remains constant for all procedures. However, altering the priorities of the data sources may change the final state in many ways. First, it may affect the employment history, e.g., the length and number of un / employment periods. Second, it may affect the selection into programs, the point in time (displacement), and participation in general. Above all, the outcome is probably influenced by the duration and point in time of un / employment periods.

Notice that Rule V0 consists of the following order: training, benefit, employment, and applicant. This is now changed in two different ways:

– The first variation leads to procedure V1 (sorting rule: training, employment, benefit, and applicant), where participation in a labor market program still has the highest priority because of the estimation at the end, which is of interest to the labor market researcher. The major difference between this and V0 is that the priority of the two sources of payments (Benefit Recipient History, Employment History) is reversed. As mentioned above, both are regarded as reliable because they include payments (benefits, wages) that must be precise. The lack of a clear indication as to which one is more accurate and therefore which one to select is a sufficient reason to analyze the impact of changing their priority. As a potential effect, changing the position of employment and benefit information in the sorting procedure gives employment information in V1 greater weight than in V0. One can therefore expect a higher number and longer durations of employment episodes in the panel data of the analysis sample.

– Procedure V2 (sorting rule: employment, training, benefit, applicant) assumes the Participants Database is not fully valid because a participant may have dropped out of the measure without correcting this in the data or a measure may have been rescheduled and both observations now occur in the data without an identifier as to which one is right. The procedure therefore downgrades the priority of this data. However, since participation can occur simultaneously with benefit receipt and since any evaluation is most interested in the effects of participation, they are not downgraded completely, that means not below benefit receipt, but categorized as second priority. Categorizing these below benefit receipt leads to a dramatic reduction of the number of participation spells available for subsequent evaluation.[8] Therefore participation in training measures is ordered below employment and above benefit receipt. Applicant Pool Data remain at the lowest level of

[8] This has been tested in a so-called “naive model”, which is not shown.

priority because no pay is involved and the state "searching for a job but not unemployed" may be parallel with nearly every other state, and no additional information is available on this. By applying this sorting rule, the employment episodes gain extra weight and observations for the treatment group may be "lost" in the control group's favor. Caution is advised in the special context of program evaluation because this relates to a problem that can be described as an increase of unobserved substitutes in the pool of the potential counterfactuals.

The consequences of the different rules for the example can be seen in Table 1. For each procedure, the row with the final states is shown. The respective order of priority is at the top of each row.

Table 1

Variation in the final states

	T1	T2	T3	T4	T5	T6	T7
V0: training, benefit, employment, applicant							
Final state	benefit	benefit	benefit	employed	training	training	training
V1: training, employment, benefit, applicant							
Final state	employed	employed	employed	employed	training	training	training
V2: employment, training, benefit, applicant							
Final state	employed	employed	employed	employed	training	training	employed

Comparing the final states of sample V1 in reference to sample V0 leads to changes in the first three timeframes. The states for the other periods remain the same. This is exactly what one would expect when reversing the priority of the Employment History and the Benefit Recipient History and may therefore have a considerable impact on the employment history before and after program participation. Comparing V2 to V0, the changes from benefit receipt to employment in the first three periods remain the same as in V1 because being employed is still of higher priority than receiving benefits. Additionally, one of the three periods of training participation changes into employment, which is in line with the expected pattern. Furthermore, dropouts from labor market programs are now taken into consideration, and the individual is employed earlier than in V0.

5. Results

5.1 Descriptive Statistics

To assess the influence of the different cleansing procedures on the real data, I compare the evaluation samples with the different underlying orders of priority. A first step in this investigation is testing differences in the sample means with reference to the benchmark sample V0. This is done for different programs. Therefore, before starting the comparisons, see Table 2, which reports a brief description of the types of programs.

The last four programs in Table 2 (JRT, GT6, GT6+, DC) are part of what is known as Further Vocational Training. In the following, the results of these programs are displayed and discussed in detail, whereas the first three programs of Table 2 (ST, SCM, JSA) belong to the group of training measures and are not displayed or discussed here[9] due to space restrictions unless they are important for the overall results (interactions). Also notice that the term "participants" in this study only covers individuals who started a program during the 18 months after becoming unemployed and received unemployment benefits immediately before starting the program.

Table 2

Description of programs

Program type	Description
Jobseeker assessment (JSA)	Assessment of jobseekers' ability and willingness to search for job and to work, basic job search assistance.
Short training (ST)	Minor adjustment of skills.
Short combined measures (SCM)	Acquisition of specific knowledge and skills
Job-related training (JRT)	Combined off-the-job and on-the-job training in a specific field of profession.
General further training $\leq$ 6 months (GT6)	General update, adjustment and extension of knowledge and skills; mainly off the job, planned duration $\leq$ 6 months
General further training $>$ 6 months (GT6+)	General update, adjustment and extension of knowledge and skills; mainly off the job, planned duration $>$ 6 months
Degree course (DC)	Vocational training that awards a formal professional degree and that corresponds to regular vocational training in the German apprenticeship system.

In Table 3 selected descriptive statistics are presented for all three samples. The selection is based on the difference between the sample and the bench-

[9] The results for these programs can be found in Appendix.

mark, and only those with a difference of greater than one percentage point are displayed.[10] The table displays the total participants and other variables for the benchmark in column two and for the two variations V1 and V2 in columns three and four.

As assumed before, the number of observations decreases in all treatment groups. Besides personal characteristics like the number of children (shift from no child to one child) or the occupational status of the last job (increasing share of clerks) the time-dependent variables show especially strong differences between the samples and for all types of programs. For example, the program start appears in samples V1 and V2 more often in 2000 than in 2002, or the time an individual is unemployed before starting a program decreases.

These differences can occur for two reasons: different compositions of the samples, or the use of another observation for the same individual with different information in case of parallelism. To examine this in Table 4, the movements of individuals between the samples V0 and V1 are displayed. See, for example, the first row of Table 4, which reports that 1,020 individuals in program ST are observed based on procedure V0. Applying procedure V1 yields 941 individuals in this program. Compared to V0, this means a loss of 79 individuals (-92 dropouts, see the last column; +14 new, see second-to-last row). However, the majority of the participants of ST in V0 (90%) are again in ST when applying V1; only one individual is now participating in JSA instead of ST and two are now in the control group (non-participants, NP).

It can be seen that a large share of all individuals (91%) participates in the same type of program in V1 as they did in V0 and that 86% of the non-participants (NP) are also not participating in a program in V1. Therefore, a change in the underlying data cleansing procedure does not lead to an overall change of the sample, and the participants who are dropping out are not moving into the group of the non-participants. The transition into other types of programs is negligible (isolated cases). The results are similar for V2.[11]

To sum up, the distinctions in the descriptive statistics occur for two reasons: sample composition (different individuals) and use of different observations (same individuals). More precisely, because the composition of the sample does change, although to a rather small extent, the differences in the mean personal characteristics can be ascribed to the dropouts and new observations that lead to the new composition. The differences in the time-dependent variables, on the other hand, do not occur due to different individuals but to changes of the final states and therefore to a prolongation or shortening of un / employment episodes. For example, the increasing number of children is very likely a result of the 10% new individuals in the sample,

10 For a full list of variables and statistics, please contact the author.

11 Table A2 in Appendix.

Table 3

Totals and shares of selected variables

Variable	Model		
	V0	V1	V2
	DC		
number of observations	503	453	447
no child	75.35	74.61	73.73
one child	13.92	15.23	15.89
completed apprenticeship	44.73	43.05	43.49
industry of last job: service	36.98	35.10	34.66
program start in 2000	20.87	22.96	22.52
program start in 2002	38.97	37.53	37.53
	GT6+		
number of observations	952	903	898
occupational status in last job: clerk	51.05	52.16	52.48
program start in 2000	24.68	25.80	25.58
time unemployed until treatment 1 – 3 months	40.23	41.31	41.46
	GT6		
number of observations	684	653	641
time unemployed until treatment 1 – 3 months	43.71	45.65	45.06
time unemployed until treatment 13 – 24 months	6.29	4.89	5.09
monthly earnings last job: 750 – 1000 EUR	28.36	26.87	27.01
	JRT		
number of observations	736	673	658
single	37.64	38.55	38.74
occupational status in last job: clerk	25.54	26.59	27.18
program start in 2000	21.47	22.60	22.67
program start in 2001	38.59	39.59	39.49
program start in 2002	39.95	37.81	37.84
remaining benefit claim >9 months	22.15	23.34	23.12
monthly earnings last job: 750 – 1000 EUR	28.67	27.62	27.63
monthly earnings last job: 1000 – 1250 EUR	18.75	19.79	19.67
time unemployed until treatment 1 – 3 months	36.01	38.40	38.44
time unemployed until treatment 7 – 12 months	27.17	26.29	26.13
time unemployed until treatment 13 – 24 months	8.97	7.83	7.96

Except the totals, all entries are in percent.

whereas the decreasing time until treatment can be traced back to a shift in the start and end dates of un / employment observations for the same individuals. These results can be confirmed with a closer look at individuals for whom the duration of employment is up to 10 months longer in V1 than in V0. On average, this difference nearly averages out to a difference of about one month.

Table 4

Transition (V0 to V1)

V0		**V1**								
	total (row)	**ST**	**SCM**	**JSA**	**JRT**	**GT6**	**GT6+**	**DC**	**NP**	**drop-outs**
ST	**1,020**	925 (90%)	0	1	0	0	0	0	2	92 (9%)
SCM	**1,252**	1	1,138 (91%)	0	1	1	0	0	0	111 (9%)
JSA	**1,415**	0	0	1,272 (90%)	1	1	0	3	3	135 (9,5%)
JRT	**736**	0	0	0	658 (89%)	0	0	1	2	75 (10%)
GT6	**684**	0	1	2	0	637 (93%)	1	0	1	42 (6%)
GT6+	**952**	0	1	1	0	3	889 (93%)	0	1	57 (6%)
DC	**503**	0	0	0	1	2	1	441 (88%)	0	58 (11,5%)
NP	**17,734**	1	3	1	3	2	3	0	15,254 (86%)	2,467 (14%)
new	645	14 (2%)	13 (2%)	20 (3%)	9 (1%)	7 (1%)	9 (1%)	7 (1%)	566 (88%)	
	total (column)	**941**	**1,156**	**1,297**	**673**	**653**	**903**	**452**	**15,829**	3,037

Note: the percentages in parentheses that relate to the rows are rounded and therefore do not necessarily need to sum to 100 over the rows.

5.2 Effect on the Estimation Results

As reported above, differences that result from the different coding decisions remain low concerning average characteristics between the sample populations. One may interpret this as a sufficient indication that outcome differences are also negligible with respect to causal effects. However, differences occur in multiple ways.

While the above investigation focuses on the composition of the comparison groups, I will now concentrate on the effect of altering the cleansing procedures on an outcome measure of interest. Based on the statistical matching framework, I focus on the potential changes in the "Average Treatment Effect on the Treated" (ATT).[12] The matching method is frequently used in evaluation studies and is also applied in the study presented by Wunsch / Lechner (2008). The matching procedure operates as a sampling device to ensure sufficient similarity between the comparisons while at the same time addressing the potential concern of having insufficient support in the comparison groups.[13]

However, when applying a matching approach, several sources of variance exist that may alter the subpopulations used for the comparison analysis. Therefore, the algorithm used to construct the comparisons is kept constant for all three models (using an Epanechnikov Kernel Matching algorithm with a fixed bandwidth based on the propensity score as the distance measure). Hence, variance is only induced by the underlying samples that are different due to the cleansing procedures applied before.

In Figure 2 the effects of participation in "job-related training" (JRT) compared to non-participation are displayed for all three models to illustrate the impact of the procedures on the estimation results[14].

For illustrative reasons, the first focus is on the general pattern of the program outcome. The reported effects are calculated on a monthly basis starting at the beginning of the treatment and show the ATT with respect to the employment status. The continuous lines show the ATT and the confidence interval for the benchmark V0. The dashed and dotted lines display the ATTs based on the samples created by the varied cleansing procedures (V1, V2). Negative values (equaling negative effects) denote worse employment chances for participants than for non-participants. Positive values, in contrast, imply better chances of being employed after having participated in a program.

What can be seen now from Figure 2 is that all procedures show almost the same pattern of ATT over time. During the first period treated individuals are locked-in the program which means they are participating and are therefore not able to be employed (month 0 – 6). This then relaxes as participants exit

12 The ATT , as referred to in the study by Wunsch / Lechner (2008), simply focuses on a specific estimator for the identification of the treatment effect.

13 Please note that this is only a very crude and brief description of the matching approach. For a more detailed description of the general setting of the matching framework, see, for example, Rosenbaum / Rubin, 1983; Rosenbaum / Rubin, 1985. More information about the exact procedure used in this study please see Heckman et al., 1998; Imbens, 2004; Caliendo / Kopeining, 2008. The program used to perform the matching is the psmatch2.ado module based on STATA 10. For more details, please contact the author.

14 See Appendix for the results of the other program types.

the program and their chances to find a job and be employed improve (recovery period; month 6 – 13). The program is pretty long and so the recovery is very slow. This implies a low ATT at the end of the observation period but with a slight upward trend. The effects of the varied procedures V1 and V2 differ slightly. V1 shows only minor differences whereas V2 has higher values, especially during the lock-in effect,[15] and thereafter is worse than or equal to the benchmark V0, but almost all differences lie within the confidence interval of V0. Although the values of V2 recover faster during the lock-in effect, the overall recovery period is not faster than for V0. To evaluate the impact of the different procedures on the results, I will take a closer look at these differences in reference to V0.

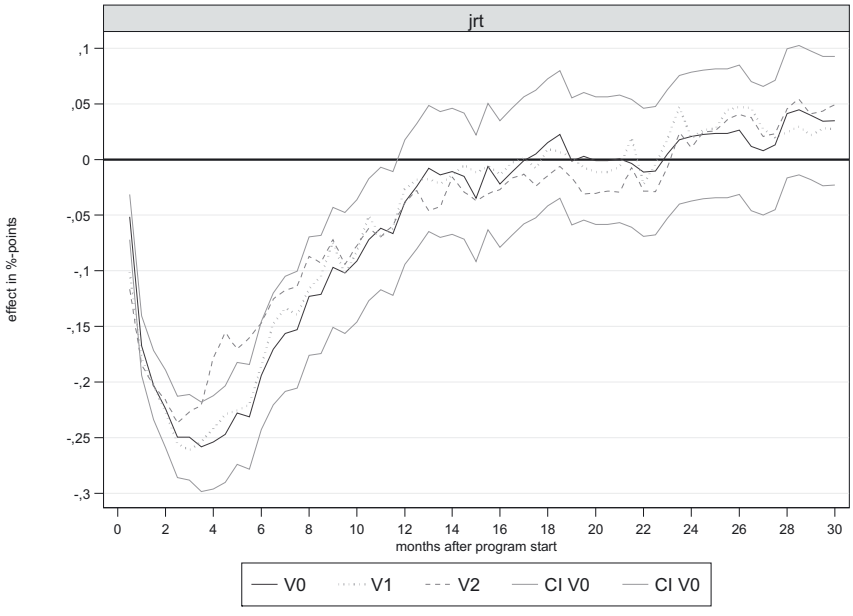


Figure 2: Effects of program participation compared to non-participation

In Figure 3, the differences between the "Average Treatment Effects on the Treated" (ATT) with respect to employment for the different procedures (V1, V2) to the benchmark V0 are displayed. They are calculated on a monthly basis starting at the beginning of the treatment period, and again, for illustrative reasons only the four programs that are part of Further Vocational Training are shown.

[15] In the terminology of van Ours (2004), lock-in effects are negative employment and earnings effects in the short run, which are directly related to program duration.

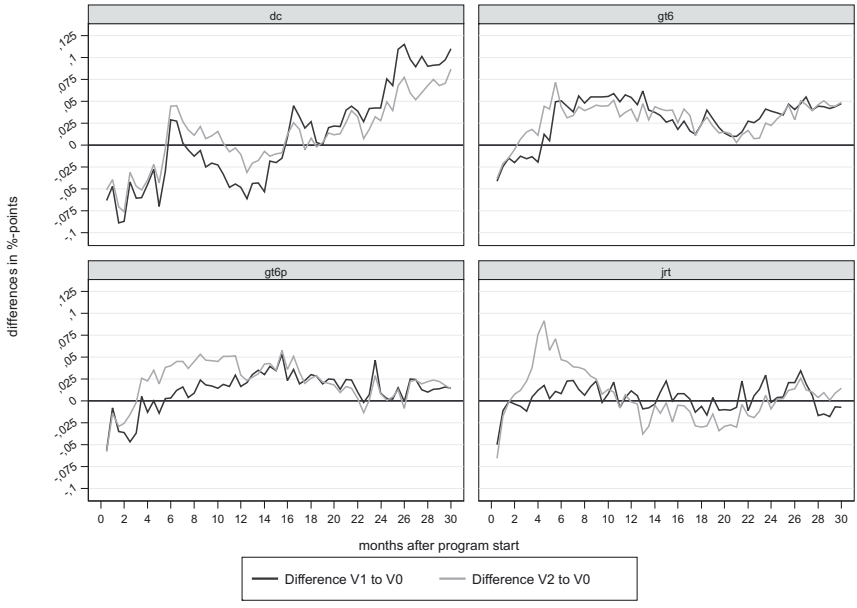


Figure 3: Differences between ATTs

As one can see, the difference between model V1 and the benchmark V0 reveals negative values in the beginning of the lock-in effect (up to –0.05) before it becomes (and stays) positive up to a maximum of 0.055 percentage points or it has an alternating sign in a smaller range (–0.0125 to 0.025). An exception to this is DC, where the range is much larger: it starts with a maximum negative value of 0.09 in the beginning of the lock-in effect, followed by a rapid increase (0.025) and a further decline before the difference to the benchmark increases considerably up to a value of 0.11 percentage points. This means that the employment chances of participants compared to non-participants in degree courses are 11% higher using model V1 – which prefers employment over benefits – than the benchmark, or they do not differ remarkably, as in JRT.

Comparing the estimation results of model V2 and benchmark V0 yields similar findings to those described for the difference V1 to V0. Deviating from these results one finds much higher differences during the lock-in effect for all types of programs. A possible explanation is the priority of the data source and therefore the change in sorting rule 2 and the selection rule. V0 grants program participation the highest priority and employment is ordered third. In V2, employment is preferred over program participation in cases of parallel information. When the program starts, the effect for V2 recovers quickly and differs from V0 to different degrees across program types, with a

maximum of 0.09 percentage points for job-related training (JRT). This could be due to participants who drop out earlier and start working. Dropouts are not always (seldom) registered, and therefore two parallel observations occur in the data. V0 continues counting this as participation, whereas in V2, employment is the final state, and therefore the lock-in effect decreases and the difference increases. Shortly after the lock-in, the values are either almost identical (GT6+, GT6) or somewhat lower (JRT, DC) than difference one. This means the employment chances are approximately 0.01 to 0.05 percentage points higher (up to 0.085 for DC) using model V2 than for the benchmark. Only for JRT are the chances lower (0.01 to 0.026%-points) between month 11 and 24 after program start.

While Figure 3 shows the time-dependent pattern of the ATT, it also depicts the cumulated effects of program participation over a certain period of time. This makes it possible to study whether low differences at single points in time may cause significant differences over time. Results are reported in Table 5. As it can easily be seen that participants face losses over the 30-month observation period in unsubsidized employment for all programs and models between two months for the shorter programs and 10 months for longer programs (DC). The differences between the models are positive but not substantial and vary between 0.07 and 0.88 months. This means that even the large differences during the lock-in effects balance out over time.

Table 5

Cumulated effects and differences (in months)

model	program			
	JRT	GT6	GT6+	DC
V0	–1.76	–2.59	–6.12	–9.82
V1	–1.70	–1.71	–5.80	–9.55
V2	–1.62	–1.71	–5.47	–9.46

6. Conclusion

The influence of variations in data cleansing on overlaps in a merged administrative data set on estimation results is a crucial issue due to the complexity of these data. Different data preparation methods might lead to different analysis samples and thus affect estimation results. This study presents different cleansing procedures and the effects of data processing that yield distinctive analysis samples and compares the descriptive and estimated program effects for participants in German labor market programs based on these samples.

In a first step, a benchmark is built using the data preparation approach applied in Wunsch / Lechner (2008) before developing two variations of the benchmark procedure in a second step and applying these by changing the priority order of the data sources. Therefore, in cases of overlapping observations, the selection rule for choosing an observation changes, and thus also the final states at these points in time. Afterwards, the influence of these different procedures on the resulting samples is tested using mean comparison tests. These tests show that there are differences in the personal and time-dependent characteristics but not to a remarkable extent, which is consistent with the findings of previous studies (e.g., Waller, 2008). The composition of the evaluation sample remains almost the same (91 %) and seems therefore to be unaffected by the cleansing procedures.

Finally, the impact of the different procedures on point estimates of matching algorithms is investigated and a sensitivity analysis is conducted. The findings emphasize the results of the mean comparison tests and differ between the types of program, over time, and across procedures. Generally the differences are substantial primarily during the lock-in effect, especially in the longer programs, and to a lesser degree at the end of the observation period. The first may be of minor importance if one is interested in long-term effects only but the latter may be of practical importance. The cumulated effects over the whole observation period balance the differences at the single points in time and do not differ to a notable extent.

Therefore the results show that data cleansing has to be done carefully and that simple, deliberate rules are necessary when it comes to overlapping observations. Not only should sensitivity analysis and robustness checks for the evaluation method be an essential part of each evaluation; the data cleansing also has to be tested if using administrative data with overlapping periods. At least two different variants of data cleansing should be tested to assess the influence on the results. A transition matrix can identify possible displacements of the sample and can reveal possible weaknesses in results, since the composition and creation of treatment and control groups is a crucial part of such evaluation methods. However, the time and efforts required to check different cleansing procedures and coding decisions should not exceed the benefits to be gained. In sum, administrative data have shown their importance and widespread applicability as a data source for empirical research. Results gained from the evaluation of administrative data are (relatively) robust to changes in the data cleansing as long as the cleansing rules are not completely beside the point.

References

Angrist, J. D. (1998): Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants, Econometrica 66(2), 249 – 288.

Ashenfelter, O. (1978): Estimating the Effect of Training Programs on Earnings, The Review of Economics and Statistics 60(1), 47 – 57.

Bauer, T. / *Bonin,* H. / *Sunde,* U. (2007): Real and Nominal Wage Rigidities and the Rate of Inflation: Evidence from German Micro Data, The Economic Journal 117, 508 – 529.

Bernhard, S. / *Dressel,* C. / *Fitzenberger,* B. / *Schnitzlein,* D. / *Stephan,* G. (2006): Überschneidungen in der IEBS: Deskriptive Auswertung und Interpretation, FDZ Methodenreport 04 / 2006, Nürnberg.

Biewen, M. / *Fitzenberger,* B. / *Osikominu,* A. / *Waller,* M. (2007): Which Program for Whom? Evidence on the Comparative Effectiveness of Public Sponsored Training Programs in Germany, IZA Discussion Paper 2885.

Blank, R. M. / *Charles,* K. K. / *Sallee,* J. M. (2009): A Cautionary Tale About the Use of Administrative Data: Evidence from Age of Marriage Laws, American Economic Journal: Applied Economics 1(2), 128 – 149.

Caliendo, M. / *Kopeinig,* S. (2008): Some Practical Guidance for the Implementation of Propensity Score Matching, Journal of Economic Surveys 22(1), 31 – 72.

Carling, K. / *Richardson,* K. (2004): The relative efficiency of labour market programs: Swedish experience from the 1990s, Labour Economics 11(3), 335 – 354.

Eliason, M. / *Storrie,* D. (2006): Lasting or Latent Scars? Swedish Evidence on Long-Term Effects of Job-Displacement, Journal of Labour Economics 24(4), 831 – 856.

Engelhardt, A. / *Oberschachtsiek,* D. / *Scioch,* P. (2008): Datengenese zweier Datenkonzepte: MTG (Maßnahme-Teilnahme-Grunddatei) und ISAAK (Instrumente Aktiver Arbeitsmarktpolitik). Eine Betrachtung ausgewählter Fälle am Beispiel der Förderung im Rahmen des ESF-BA-Programms, FDZ Methodenreport 08 / 2008, Nürnberg.

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

Fitzenberger, B. / *Osikominu,* A. / *Völter,* R. (2009): Get Training or Wait? Long-Run Employment Effects of Training Programs for the Unemployment in West Germany, Annales d'Economie et de Statistique, (forthcoming).

Fitzenberger, B. / *Speckesser,* S. (2007): Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany, Empirical Economics 32(2), 529 – 573.

Fitzenberger, B. / *Wilke,* R. A. (2009): Unemployment Durations in West-Germany Before and After the Reform of the Unemployment Compensation System during the 1980s, German Economic Review, (forthcoming).

Geerdsen, L.P. / *Holm,* A. (2004): Job-search Incentives from Labor Market Programs – an Empirical Analysis, Centre for Applied Microeconometrics Working Paper 2004-03.

Groves, R. M. (2004): Survey Errors and Survey Costs, Hoboken, NJ.

Groves, R. M. / *Fowler,* F. J. / *Couper,* M. P. / *Lepkowski,* J. M. / *Singer,* E. / *Tourangeau,* R. (2004), Survey Methodology, Hoboken, NJ.

Hämäläinen, K. / *Ollikainen,* V. (2004): Differential Effects of Active Labour Market Programmes in the early stages of young people's unemployment, VATT Research Reports 115.

Heckman, J. J. / *Ichimura,* H. / *Todd,* P. E. (1998): Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, The Review of Economic Studies 65(2), 261 – 294.

Hotz, V. J. / *Scholz,* J. K. (2001): Measuring Employment and Income for Low-Income Populations with Administrative and Survey Data, IRP Discussion Paper, 1224-01.

Hujer, R. / *Caliendo,* M. / *Thomsen,* S. L. (2004): New evidence on the effects of job creation schemes in Germany – a matching approach with threefold heterogeneity, Research in Economics 58(4), 257 – 302.

Imbens, G. W. (2004): Nonparametric Estimation of Average Treatment Effects under Exogeneity: a Review, The Review of Economics and Statistics 86(1), 4 – 29.

Jabine, T. B. / *Scheuren,* F. (1985): Goals for Statistical Uses of Administrative Records: The Next 10 Years, Journal of Business & Economic Statistics 3(4), 380 – 391.

Jacobebbinghaus, P. / *Seth,* S. (2007): The German integrated employment biographies sample IEBS, Schmollers Jahrbuch 127(2), 335 – 342.

Jaenichen, U. / *Kruppe,* T. / *Stephan,* G. / *Ullrich,* B. / *Wießner,* F. (2005): You can split it if you really want: Korrekturvorschläge für ausgewählte Inkonsistenzen in IEB und MTG, FDZ Datenreport 04 / 2005, Nürnberg.

Jenkins, S. P. / *Lynn,* P. / *Jäckle,* A. / *Sala,* E. (2004): Linking Household Survey and Administrative Record Data: What Should the Matching Variables Be?, DIW Discussion Papers 489, Berlin.

Johansson, P. / *Skedinger,* P. (2005): Misreporting in register data on disability status: evidence from the Swedish Public Employment Service, Empirical Economics 39(2), 411 – 434.

Klose, C. / *Bender,* S. (2000): Berufliche Weiterbildung für Arbeitslose – ein Weg zurück in Beschäftigung? Analyse einer Abgängerkohorte des Jahres 1986 aus Maßnahmen zur Fortbildung und Umschulung mit einer ergänzten IAB-Beschäftigungsstichprobe 1975 – 1990, Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 24, 421 – 444.

Kluve, J. / *Card,* D. / *Fertig,* M. / *Gora,* M. / *Jacobi,* L. / *Jensen,* P. / *Leetmaa,* R. / *Nima,* L. / *Patacchini,* E. / *Schaffner,* S. / *Schmidt,* C. M. / *van der Klaauw,* B. / *Weber,* A. (2006): Active Labor Market Policy in Europe: Performance and Perspectives, Berlin / Heidelberg.

Köhler, M. / *Thomsen,* U. (2009): Data integration and consolidation of administrative data from various sources – the case of Germans' employment histories. Historical Social Research 34(3), 215 – 229.

Kruppe, T. / *Müller,* E. / *Wichert,* L. L. / *Wilke,* R. A. (2008): On the Definition of Unemployment and its Implementation in Register Data – The Case of Germany, Schmollers Jahrbuch 128(3), 461 – 488.

Kruppe, T. / *Oertel,* M. (2003): Von Verwaltungsdaten zu Forschungsdaten – Die Individualdaten für die Evaluation des ESF-BA-Programms 2000 – 2006, IAB Werkstattbericht, Nürnberg.

KVI: Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (eds.) (2001): Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik, Baden-Baden (mit CD-ROM mit 41 Expertisen und Beiträgen, die im Auftrag der Kommission bzw. im Zuge der Kommissionsarbeiten erstellt wurden).

Lechner, M. / *Miquel,* R. (2009): Identification of the effects of dynamic treatments by sequential conditional independence assumptions, Empirical Economics, online first.

Lechner, M. / *Miquel,* R. / *Wunsch,* C. (2004): Long-Run Effects of Public Sector Sponsored Training in West Germany, IZA Discussion Paper, 1443.

Mueser, P. R. / *Troske,* K. R. / *Gorislavsky,* A. (2007): Using State Administrative Data to measure Program Performance, The Review of Economics and Statistics 89(4), 761 – 783.

Oberschachtsiek, D. / *Scioch,* P. (2009): Cleansing Procedures for Overlaps and Inconsistencies in Administrative Data. The Case of German Labour Market Data, Historical Social Research 34 (3), 242 – 259.

Osikominu, A. (2009): Quick Job Entry or Long-Term Human Capital Development? The Dynamic Effects of Alternative Training Schemes, IZA Discussion Paper 4638.

Pyy-Martikainen, M. / *Rendtel,* U. (2008): Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys, AStA 92, 297 – 318.

Reimer, M. / *Künster,* R. (2004): Linking job episodes from retrospective surveys and social security data: specific challenges, feasibility and quality of outcome, Max-Planck-Institut für Bildungsforschung, Berlin (Arbeitsberichte Max-Planck-Institut für Bildungsforschung, Forschungsbereich Bildung, Arbeit und gesellschaftliche Entwicklung 2004, 8).

Rendtel, U. / *Nordberg,* L. / *Jäntti,* M. / *Hanisch,* J. U. / *Basic,* E. (2004): Report on quality of income data, CHINTEX Working Paper 21.

Rinne, U. / *Uhlendorff,* A. / *Zhao,* Z. (2008): Vouchers and Caseworkers in Public Training Programs: Evidence from the Hartz Reform in Germany, IZA Discussion Papers 3910.

Røed, K. / *Raaum,* O. (2003): Administrative Registers – Unexplored Reservoirs of Scientific Knowledge?, The Economic Journal 113, 258 – 281.

Rosenbaum, P. R. / *Rubin,* D. B. (1985): Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, The American Statistician 39(1), 33 – 38.

Rosenbaum, P. R. / *Rubin,* D. B. (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrica, 70 (1), 41 – 40.

Seysen, C. (2009): Effects of changes in data collection mode on data quality in administrative data – the case of participation in programmes offered by the German Employment Agency, Historical Social Research 34(3), 191 – 203.

Stephan, G. (2008): The effects of active labor market programs in Germany – an investigation using different definitions of non-treatment, Jahrbücher für Nationalökonomie und Statistik 228(5 / 6), 586 – 611.

van den Berg, G. J. / *Lindeboom,* M. / *Dolton,* P. J. (2004): Survey Non-Response and Unemployment Duration, IZA Discussion Papers 1303.

van Ours, J. (2004): The Locking-in Effect of Subsidized Jobs, Journal of Comparative Economics 32, 37 – 52.

Waller, M. (2008): On the Importance of Correcting Reported End Dates of Labor Market Programs, Schmollers Jahrbuch 128, 213 – 236.

Wallgren, A. / *Wallgren,* B. (2007): Register-based Statistics – Administrative Data for Statistical Purposes. Hoboken, NJ.

Wunsch, C. / *Lechner,* M. (2008): What did all the money do? On the General Ineffectiveness of Recent West German Labour Market Programmes, Kyklos 61(1), 134 – 174.

Appendix

Table A1

Descriptive results for training programs

Variable	Model		
	V0	**V1**	**V2**
	ST		
number of observations	1,020	941	917
female	49.61	47.61	48.06
no child	60.78	62.38	62.15
one child	19.71	17.96	18.17
looking for fulltime-job only	77.75	78.85	79.35
occupational status in last job: clerk	42.65	43.46	43.76
program start in 2002	12.06	11.26	10.97
time unemployed until treatment 1 – 3 months	43.92	42.72	42.69
	SCM		
number of observations	1,252	1,156	1,118
looking for fulltime-job only	76.68	77.42	77.74
last occupation: services	37.54	36.33	36.66
program start in 2002	39.14	38.06	38.52
time unemployed until treatment 1 – 3 months	35.62	36.85	36.57
time unemployed until treatment 10 – 12 months	28.83	27.68	28.18

Continued next page

(Continued Table A1)

Variable	Model		
	V0	V1	V2
	JSA		
number of observations	1,415	1,297	1,249
qualification in desired job: skilled	42.69	43.41	43.85
monthly earnings last job: 500 – 750 EUR	25.72	24.21)	24.37
time unemployed until treatment 1 – 3 months	40.14	42.95	42.35
time unemployed until treatment 10 – 12 months	25.23	22.90	22.95
	NP		
number of observations	17.734	15,829	15,276
last occupation: services	36.32	34.88	34.73
time unemployed until treatment 4 – 6 months	58.85	62.41	61.49
time unemployed until treatment 10 – 12 months	31.10	27.20	28.00

All entries are in percentages. Differences to V0 are displayed in percentage points in parentheses.

Table A2

Transition (V0 to V2)

V0		V1								
	total (row)	**ST**	**SCM**	**JSA**	**JRT**	**GT6**	**GT6+**	**DC**	**NP**	**drop-outs**
ST	**1,020**	906 (89%)	0	1	0	0	0	0	7	106 (10%)
SCM	**1,252**	2	1,107 (88%)	0	1	1	0	0	5	136 (11%)
JSA	**1,415**	2	1	1,238 (87%)	0	0	0	3	11	160 (11%)
JRT	**736**	0	0	1	650 (88%)	0	0	3	3	80 (11%)
GT6	**684**	0	1	1	1	629 (92%)	1	0	1	50 (7%)
GT6+	**952**	0	1	0	1	2	886 (93%)	0	3	59 (6%)
DC	**503**	0	0	0	0	1	1	434 (86%)	1	66 (13%)
NP	**17,734**	1	2	0	4	3	3	0	15,134 (85%)	2,587 (15%)
new	151	6 (4%)	6 (4%)	8 (5%)	1 (1%)	5 (3%)	7 (5%)	7 (1%)	111 (73%)	
	total (column)	**917**	**1,118**	**1,249**	**658**	**641**	**898**	**447**	**15,276**	3,244

Note: the percentages in parentheses are rounded and therefore do not necessarily need to sum to 100.

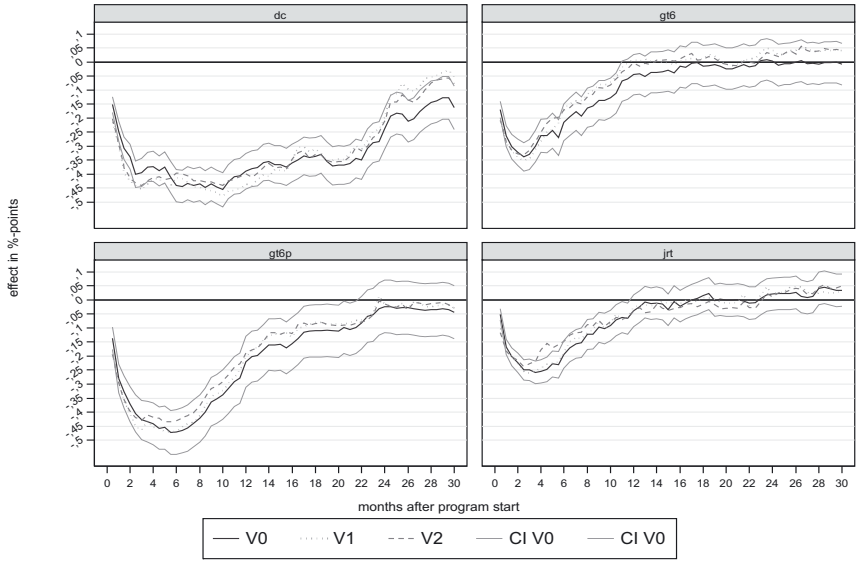


Figure A1: Effects of program participation compared to non-participation (Further Vocational Training)

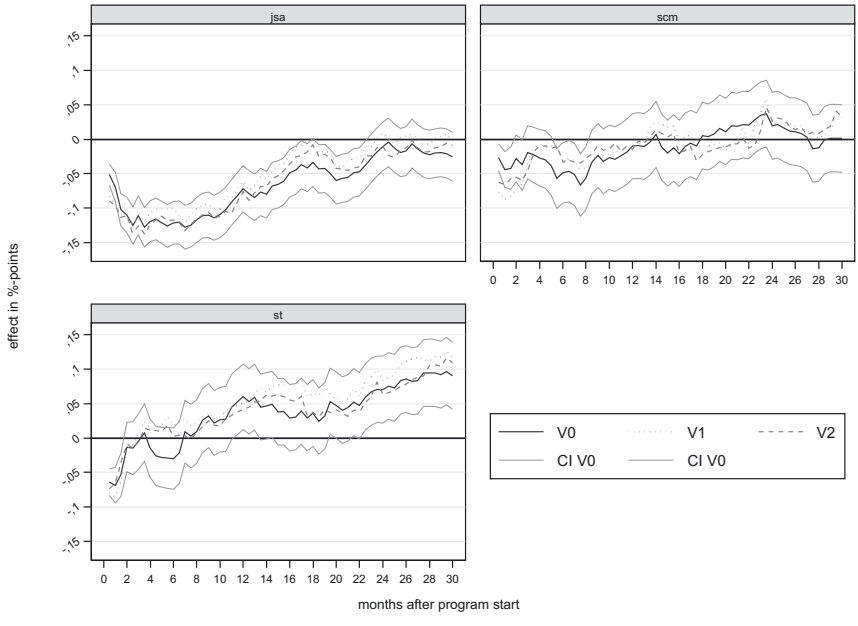


Figure A2: Effects of program participation compared to non-participation (Training Programs)

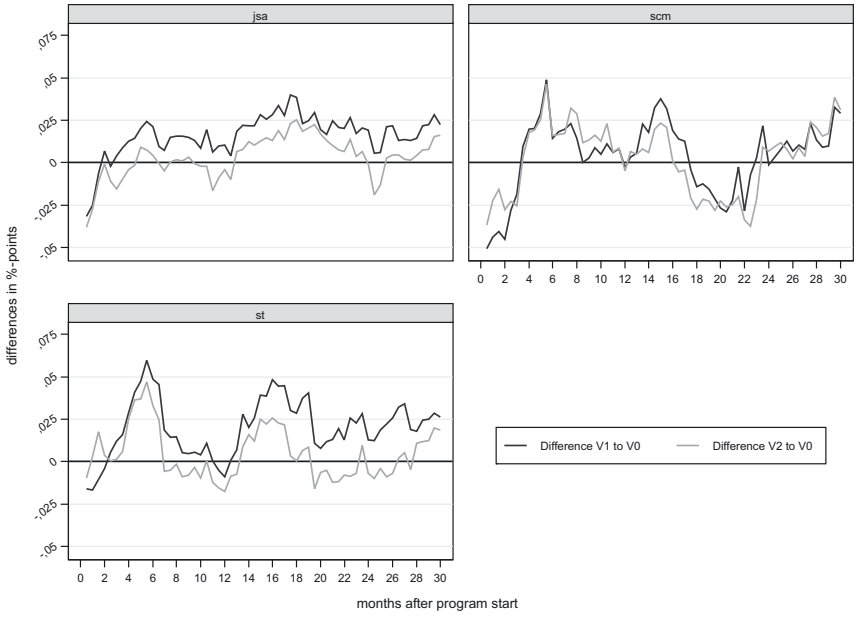


Figure A3: Differences ot the ATTs (Training Programs)

Table A3

Cumulated effects and differences (training programs)

model	program		
	ST	SCM	JSA
V0	1.10	–0.31	–1.98
V1	1.70	–0.25	–1.49
V2	1.24	–0.28	–1.88