Lechner, Michael; Wunsch, Conny

Conference Paper

Which control variables do we really need for matching based evaluations of labour market programmes?

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Evaluation Econometrics, No. A6-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Lechner, Michael; Wunsch, Conny (2010) : Which control variables do we really need for matching based evaluations of labour market programmes?, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Evaluation Econometrics, No. A6-V1, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37434

Which control variables do we really need for matching based evaluations of labour market programmes?

Michael Lechner and Conny Wunsch*

Swiss Institute for Empirical Economic Research

University of St.Gallen

This draft: January 2010

Date this version has been printed: *28 January 2010*

Incomplete draft

Please do not quote nor circulate without permission of one of the authors

Comments are very welcome

Abstract: Based on a new, exceptionally informative and large German linked employer-employee administrative dataset, we investigate the question whether the omission of important control variables in matching estimation leads to biased impact estimates of typical active labour market programmes for the unemployed. Such biases would lead to false policy conclusions about the effectiveness of these expensive policies. Based on our preliminary findings, it seems that controlling for standard demographic variables removes already a substantial part of the potential selection bias.

Keywords: Training, job search assistance, matching estimation, active labour market policies

JEL classification: J68

Address for correspondence: Michael Lechner, Conny Wunsch,
Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, Switzerland, Michael.Lechner@unisg.ch, Conny.Wunsch@unisg.ch, www.sew.unisg.ch/lechner.

---

* The first author is also affiliated with ZEW, Mannheim, CEPR and PSI, London, IZA, Bonn, and IAB, Nuremberg. The second author is also affiliated with CESifo, Munich. This project received financial support from the Institut für Arbeitsmarkt und Berufsforschung, IAB, Nuremberg, and from the St. Gallen Research Center in Aging, Welfare, and Labour Market Analysis (SCALA). Parts of the paper were written while the first author visited CESifo, Munich, and the second author visited IFS/UCL, London. The hospitality and support of both institutions is gratefully acknowledged. The usual disclaimer applies.

1. Introduction

Costing up to 3% of GDP (OECD, 2008) active labour market policies aiming at bringing the unemployed back to work belong to the most important public expenditure programmes in OECD countries. This has led to considerable and increasing interest among both policy makers and researchers to quantify the effects of participating in these programmes on the labour market outcomes of their participants.

Well designed and implemented social experiments where jobseekers are randomly assigned to programmes are probably the most convincing way to assess the performance of such programmes. However, they are very costly and often suffer from small sample size, implementation problems and a lack of representativeness (e.g., Heckman and Smith, 1995). Moreover, in Europe they are usually not implementable because of the argument that helpful services cannot be denied arbitrarily. In the absence of random assignment there is the problem that confounding variables lead to spurious correlations of the various outcome variables with the variables indicating programme participation.

An increasing number of evaluation studies, especially for Europe, argue that the data they use are informative enough to capture all potential confounders and thus allow for matching estimation.[1] Of course, the key assumption in these studies is that the data contain all variables that jointly influence outcomes and programme participation. If this assumption is true, controlling for these confounding variables will solve the spurious correlation issue by removing selection bias and identify average programme effects with a minimum of further assumptions required. The literature in this field is very advanced with respect to the

[1] Among the many studies, see for example Dorsett (2006) for the UK, Larsson (2003) and Sianesi (2004) for Sweden, Gerfin and Lechner (2002) for Switzerland, Lechner, Miquel and Wunsch (2010) for Germany, Jespersen, Munch and Skipper (2008) for Denmark and Heinrich et al., (2009) for the U.S.

econometric methods and many benchmark applications exist. The methodological advances are nicely summarized in the comprehensive surveys by Blundell and Costa-Dias (2009) and Imbens and Wooldridge (2009). The large recent applied literature is for example covered in the meta-study by Card, Kluve, and Weber (2009).

Many governments, especially in Europe but increasingly also in the US, have become aware of the value of informative and accurate data to perform cost-benefit analyses and obtain reliable impact estimates. Thus, they are prepared to give the scientific community access to rich administrative data bases that foster such evaluation studies. Moreover, despite the availability of good data, the recent advances in methods, and the many applications that appeared recently, and despite every involved researcher agreeing that the plausibility of the so-called 'selection-on-observables' assumption is the key for the credibility and policy relevance of the results, there is not yet any systematic investigation of exactly which variables are required such that this assumption can be safely assumed (i.e. without incurring much selection bias).

One potential reason for this gap in the extensive literature on the evaluation of labour market programmes by matching methods is the missing reference case. Such a reference case would reveal the true effect of the programme and then allow a judgement of the bias that would occur for specific sets of control variables. In fact, a social experiment might be a good reference case if it is well-conducted, has a large enough sample to precisely determine the 'truth', and is representative for the programmes at issue. For some US programmes, such benchmark studies exist (although they may in many cases not meet all these requirements).[2] For reasons outlined above, no such experimental benchmarks exist for European pro-

[2] See, for example LaLonde (1986), Heckman and Hotz (1989), Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1999), Smith and Todd (2005), Dehejia (2005), Peikes, Moreno, and Orzol (2008), Shadish, Clark, and Steiner (2008).

grammes. The only results based on quasi-experiments, for example like instrumental variable estimation (see for example Frölich and Lechner, 2010) or difference-in-difference estimation (see for example Petrongolo, 2009), have the drawback that they identify different, instrument specific parameters. This makes any cross-study comparison obviously difficult.

It is goal of this paper to fill part of this gap in the literature and to better understand which variables are most likely required as control variables for classical evaluation studies of typical active labour market programmes. Below, we argue that new German administrative data is informative enough to provide a credible benchmark study. Of course, one may always hope to be able to conduct experiments that cover all important active labour market programmes (and then wait a couple of years to observe the relevant medium term outcomes), this seems unlikely to happen in the near future. Thus, creating a benchmark from an observational study with exceptionally rich data seems to be the only way possible to understand the important issues at stake, namely the relevance of different types of control variables for the policy relevance of the programme effects.

Besides providing a credible benchmark for just one country, the value of the analysis is clearly enhanced if it is also informative and relevant for programmes and evaluation studies outside Germany that use potentially different data. We argue that our data covers all types of variables typically included in the various published evaluation studies so that the effect of lacking a particular type of information on the estimates of the programme selection process as well as on the effects of the programme can be simulated. This is incidentally the main reason why (besides other potential problems mentioned above) the social experiments conducted in the U.S. and corresponding U.S. data cannot be used for such an exercise: As we are going to argue below, the U.S. data lack important information that is available in European datasets. We also argue that the German programmes we analyse, namely job search assistance and training, are not only the most widely used programmes in OECD countries but also

typical in terms of their contents, implementation, and selection of participants. Furthermore, our data allow constructing the set of outcome variables that is typically used in evaluation studies. Finally, to focus on a typical Western-style developed economy we restrict the analysis to West-German programmes.

Our preliminary results surprisingly suggest that indeed already a limited set of demographic variables removes a considerable part of the selection bias. These findings will be extended soon when the still missing results are available.

The remainder of the paper is organized as follows. The following section describes in detail the programmes that will be analysed, how they compare to programmes used in other countries and how selection of participants works. In Section 3, we provide all details on the data, how they relate to other data available, and why they justify identification of programme effects based on a selection on observables (conditional independence) assumption. We also describe our matching procedure. Section 4 analyses the selection into the programmes based on all available confounders and analyse the relevance of the different blocs of confounding variables. It also proves the benchmark programme effects. Section 5 analyses the changes in the estimates compared to the benchmark that occur when fewer variables are available. It also relates these changes to changes in the propensity and relevance of the confounders for various outcome variables considered. The last section concludes. Two appendices as well as an additional Internet Appendix contain further details on the data and the estimation.

2. The determinants of participation in typical labour market programmes

2.1 Programmes considered

In order to allow drawing conclusions that are relevant for a large part of the field, we focus the analysis on the two types of active labour market programmes for the unemployed

that are most widely used in Western-style developed economies: job search assistance and vocational training for skill-upgrading.

The type of job search assistance programme implemented in Germany is very representative for this class of programmes (e.g., Thomsen, 2009). It comprises the typical combination of counselling services, referral to vacancies, monitoring in the form of availability checks, one-day trial internships of potential candidates in firms for specific vacancies, and job search training where jobseekers learn how to locate job vacancies, how to write an application and where they practice job interviews.

German training programmes include those types of programmes commonly used in most other OECD countries,[3] but the overall range of training programmes is more diverse with respect not only to the form and intensity of the human capital investment involved but also to their respective duration (ranging from several weeks to more than two years). We restrict our analysis to the internationally most typical training programmes which comprise occupational skills training, skill upgrading and programs that combine workplace training with related instruction, and that have planned durations of no more than six months.

The implementation of the two types of German programmes we look at is also largely representative with respect to eligibility and selection into the programmes. Job search assistance is used relatively early in the unemployment spell and for a rather wide range of types of unemployed. Training starts somewhat later in the unemployment spell and courses are targeted more specifically towards unemployed with certain qualification needs. In the period we consider, 2000-2002, eligibility for programme participation required jobseekers to qualify

[3] Before the Workforce Investment Act (WIA) became effective in the U.S. in 2000, the German programmes where only representative for European programmes, because the U.S. Job Training Partnership Act (JTPA) programmes used before WIA focused mainly of pre-vocational as well as literacy and English as a foreign language training. With the WIA a range of training programmes has been introduced in the U.S. that is very similar to the European programmes.

for unemployment insurance (UI) payments (so-called unemployment benefits), or for unemployment assistance which was a means-tested benefit that was paid after exhaustion of UI benefits from tax revenue. See Wunsch and Lechner (2008) for a detailed description of the scope and volume of the German programmes and their participants in the period we consider here (2000-2002).

2.2 Participation in the programmes

In general, programme participation is the outcome of decisions made by both the caseworker and the unemployed person. Usually the caseworker proposes participation in a programme to improve a client's employment prospects, though sometimes the jobseeker also proposes a programme. In either case, the jobseeker must apply for permission before beginning any subsidised programme. The caseworker decides whether or not the applicant will be admitted. There is no legal entitlement to participation, and caseworkers have a considerable amount of discretion. Normally the caseworker decides in consultation with the potential participant whether or not to enter a programme and, if so, what kind would be appropriate based on an assessment of the jobseeker's employment prospects and the specific qualification needs of the unemployed. According to the German legislation, they also have to take into account the chances of successful completion of the programme, and the local labour market conditions. Similar arguments apply to self-selection by the unemployed because they most likely compare their employment prospects with and without a specific programme, as well as the corresponding costs in terms of effort and alternative ways to spend their time, or potentially foregone benefits in case of refusal to participate.

Similar to many other countries there are also institutional incentives to participate in labour market programmes. Jobseekers who refuse to participate in a programme they have been assigned to by the caseworker risk a benefit sanction, i.e. a temporary cut or suspension of their unemployment benefit or unemployment assistance. Moreover, and this is a feature

mainly of some European countries, participation in training programmes stops the clock for UI claims during participation, meaning that the UI claim at the beginning of training is the same as at the end of training.[4] Thus, jobseekers can effectively extend their UI claim by participating in a programme. This is, however, not true for job search assistance, where the unemployed, if eligible, continue to receive their UI benefit and potentially use up their UI claim.

The implications of the described selection process for strategies to identify the causal effects of job search assistance and training programmes on labour market outcomes that are based on selection on observables can be summarized as follows: First, it is important to note that all determinants of programme participation mentioned above are likely to affect labour market outcomes like employment status and earnings, making them potential confounders. Therefore, we discuss all of them in turn with respect to the required measurements they imply if we want to control for these factors in an empirical analysis.

To ensure eligibility for programme participation, we have to determine whether unemployed individuals qualify for unemployment benefits or assistance. Moreover, to capture institutional incentives we must observe the amount of the benefits, UI eligibility and the remaining UI claim. Next, we need to be able to capture the main determinants of employment prospects, which include individual characteristics like age, gender, marital status, presence of (young) kids, education, skills, productivity, motivation as well as work, occupation and industry-specific experience but also local labour market conditions. According to the German legislation, the latter also have a direct impact on the participation decision. To determine qualification needs we must also capture education, skills and the different types of

[4] In the 1990s, participation in training even counted towards acquisition of new UI claims. Since 2005, UI claims are reduced by half of the duration of training.

work experience, as well as what kind of job a person is looking for in order to determine the required target skills. Moreover, for job search assistance, it is also relevant whether the job-seeker has previous unemployment experience that makes him familiar with job search or whether he comes from a declining industry/occupation that may require him to look for jobs in other industries/occupations where he may be inexperienced. The latter is also relevant for potential training needs. For the probability of successful programme completion essentially the same factors play a role as for employment prospects and qualification needs. The final set of factors is related to preferences and alternative ways of using the time out of employment. The most relevant cases are women's fertility decisions, the main determinants of which would have to be captured. In particular, Lechner and Wiehler (2010) show that for females programme participation and becoming pregnant during unemployment are both attractive options. For men alternative time use is largely negligible because institutions provide strong incentives to leave unemployment if the opportunity occurs.

3. Data and econometric methodology

3.1 Data

We use a unique linked employer-employee administrative data base. It is probably *the* most informative data base that is currently available for evaluating typical labour market programmes (see Section 3.5 for a discussion of how our data compare to other available data bases). The data comprise a 2% random sample drawn from the population of all German employees subject to social insurance[5] since 1990. They cover the period 1990-2006 and combine information from different administrative sources: (1) the records provided by employers

[5] This covers 85% of the German workforce. It excludes the self-employed as well as civil servants.

to the social insurance for each employee (1990-2006), (2) the unemployment insurance records (1990-2006), (3) the programme participation register of the Public Employment Service (PES, 2000-2006) as well as (4) the jobseeker register of the PES (2000-2006). Because these records are used to determine social insurance and unemployment benefit claims as well as programme eligibility, the data are very accurate with respect to employment status, earnings from employment, amount and duration of UI claims, and programme participation status. Moreover, the information collected by the PES on jobseekers is good as well, because it is used for counselling, job referral, monitoring, and assessing jobseeker's compliance with job search requirements.

Whenever an individual in our sample appears in one of the four registers in the period 1990-2006, we observe the corresponding spell with all available covariates. Moreover, whenever a person is employed we observe the corresponding employer information. They comprise the size, age and industry of the firm, and the composition of its workforce in terms of gender, nationality, age, education, work hours, earnings, tenure, turnover, and occupations. The latter variables are calculated from (1) from the population of all employees of the firm as of June 30 of each year 1990-2006 where the firm existed (so-called establishment history panel or *Betriebshistorikpanel*, BHP). Finally, a variety of regional information has been matched to the data via the official codes of the 439 German districts (*Kreiskennziffer*). It contains population density, migration and commuting streams, average earnings, GDP growth, unemployment rate, long-term unemployment, welfare dependency rates, urbanisation, child care and public transport facilities.

For each individual the data comprise all aspects of their employment, earnings and UI history since 1990 including day of beginning and end of each spell, type of employment (full/part-time, high/low-skilled), occupation, earnings, type and amount of UI benefit, remaining UI claim, compliance with benefit conditions (e.g. failure to show up at interview,

refusal to participate in assigned labour market program, imposition of sanction), and period when a UI recipient has reported in sick to the UI. Moreover, they cover all spells of participation in the major German labour market programmes from 2000 onwards with exact beginning, end and type of programme as well as the planned end date for the training programmes. The jobseeker register contains a wealth of individual characteristics, including date of birth, gender, educational attainment, marital status, number of kids, age of youngest child, nationality, profession, the presence of health impairments and disability status. With respect to job search the data contain the type of job looked for (full/part-time, high/low-skilled, occupation), whether the jobseeker is fully mobile within Germany and whether she has health impairments that affect employability. Moreover, the data record how many job referrals the jobseeker got from the PES, i.e. proposals by the caseworker to apply for a specific vacancy.

3.2 Sample selection and definition of participation status

Since we are interested in evaluating typical labour market programmes in industrialized economies, we restrict the analysis to the territory of former West Germany (without Berlin). We start from a sample that covers all entries into unemployment in the period 2000-2002. We exclude unemployment entries in January-March 2000 because with programme information starting only in January 2000 we want to make sure that we do not accidentally classify entries from employment programmes (which we would consider as unemployed) as entries from unsubsidized employment because the accompanying programme spell is missing. Furthermore, we restrict the analysis to the prime-age population aged 20-59 in order to avoid having to model educational choices or (early) retirement decisions. We also require individuals to be eligible for programme participation by imposing that individuals must qualify for unemployment benefits or unemployment assistance. Finally, we exclude the few cases

that start unemployment with a programme or have no information from the jobseeker register.

As in Lechner, Miquel and Wunsch (2010) and Lechner and Wunsch (2009) we define as (non-) participants all those individuals in our sample who (do not) start a programme within the first 12 months of unemployment.[6] Of course, related to the arguments of Fredriksson and Johansson (2003, 2008) and Sianesi (2004) this raises issues about dynamic programme assignment and future labour market outcomes of the so-defined non-participants. However, as long as we condition on time to treatment, it does not affect our ability to model selection into the programmes given the data. Moreover, we are only interested in comparing different models for selection correction and all specifications will be based on the same treatment definition. To focus on the internationally most widely used types of programmes, we only consider participants whose first programme is job search assistance, or training with a planned duration of no more than six months. The latter restriction is imposed in order to focus on a typical training programme, as German programmes are sometimes unusually long compared to other countries.

In order to determine time to treatment and to measure outcomes relative to programme start we simulate hypothetical programme start dates for non-participants by drawing randomly from the empirical distribution of start dates of programme participants. We do not employ approaches that condition on covariates in order to prevent any type of selection correction at this stage. The simulation is done separately for job search assistance and training because they show rather different distributions of start dates.[7] This implies that we have different samples of non-participants for job search assistance and training. We then impose

[6] Note that non-participation means not starting any programme, not just the programme used for the particular comparison.

[7] Job search assistances is used very early in the spell, while training starts later.

hypothetical programme eligibility on non-participants by requiring them to be unemployed and eligible for unemployment benefits or assistance at simulated programme start. Moreover, we discard all actual and hypothetical programme starts after 2002 to ensure that outcomes can be observed for up to four years after programme start.

3.3 Credibility of matching: Do we observe all relevant factors in this study?

At the end of Section 2 we summarized all factors that should be controlled for when identifying causal effects of the two programmes on labour market outcomes based on an selection-on-observables approach. Here we relate them to the available data and discuss the topics in turn: *Eligibility* for programme participation is ensured by the construction of the sample. Concerning the *institutional incentives* we directly observe amount of benefits, UI eligibility and remaining UI claim. To measure *local labour market conditions* we observe a rich set of regional indicators listed in Section 3.1 that allow controlling for the relevant regional differences in a detailed way.

The *determinants of employment prospects* are captured by personal characteristics like age, gender, marital status, nationality, number of kids, and age of youngest child. Furthermore, skills are measured in terms of schooling and vocational training as well as the skill profile of the last job held. Productivity is approximated by the earnings from the last job (controlling for full/part-time) as well as average earnings from employment in the last 10 years before unemployment. In addition, we observe several variables indicating health problems and whether these affect employability. Work, occupation and industry-specific experience can be calculated from 10 years of pre-unemployment employment histories and the corresponding firm data. Finally, unobserved heterogeneity in motivation, productivity and employability is captured indirectly in several ways: First, by using the 10 years of detailed employment histories to control for the quality and stability of employment, frequency and duration of previous unemployment experience, as well as other periods of non-employment;

second, by conditioning on the characteristics of the last employer that may reveal specific types of workers; third, by controlling for incidence of non-compliance with benefit conditions during past unemployment spells; and finally, by accounting for the average number of job referrals by the PES per day. This measure summarizes both the demand for the particular skill mix of the jobseeker, and the caseworkers' personal judgement of the employability of the worker. Moreover, we know whether the jobseeker is fully mobile within Germany.

In addition to the factors like skills, productivity, experience and motivation already mentioned, to proxy for the *determinants of qualification needs* we are able to account for the type of job looked for in terms of full/part-time, high/low-skilled and occupation to determine potential qualification needs. Moreover, taking up the discussion from Section 2.2 about the need to change industry or occupation we also know from which industries and occupations jobseekers come. Finally, we can capture potential job search experience and job search skills by past unemployment experience and their average duration.

Preferences for leisure and the determinants of fertility decisions of females are, of course, unobserved. However, we can capture them indirectly to the extent to which they have impacted on the employment history in the 10 years preceding unemployment. In particular, we observe the incidence and duration of unemployment as well as other forms of non-employment. Note that the latter, in addition to the number of kids and the age of the youngest child is likely to capture some aspects of fertility decisions and child raising preferences.

In summary, with the exception of some aspects of preferences our unique data enables us to capture all important confounding factors that affect both programme participation and labour market outcomes. Table 3.1 summarizes the blocks of variables that we use to control for selection. Moreover, because of the relevance of female preferences regarding fertility and child raising but limited information to capture these with our data, we are more confident regarding our ability to correct for selection for males.

Table 3.1: Control variables

No.	Block	Variables
0	Baseline characteristics	Age, gender, school degree, vocational degree, nationality, number of kids, age of youngest child, marital status, region (state dummies)
1	Timing of entry into unemployment and programme	Fortnight and quarter of entry into unemployment, time to treatment, several interaction terms
2	Last employment: non-firm characteristics	Earnings, skill profile, full/part-time, occupation
3	Last employment: firm characteristics	Firm age, number of employees, closed firm, fraction low-income, temporary and part-time jobs, age distribution, mean tenure, fraction of jobs destroyed, industry
4	Short-term employment history (up to 2 years before unemployment)	Fortnights employed/unemployed in 6/12/24 months before, fortnights out of labour force in the 6/24 months before, in programme in the 6/24 months before, no employment/unemployment in last 2 years, time since last employment/unemployment in last 2 years, unemployed/out of labour force in month 6/24 before
5	Long-term employment history (up to 10 years before unemployment)	Fortnights employed/unemployed/out of labour force in the last 4/10 years before, in programme in the last 4/10 years before, unemployed/out of labour force in month 48 before, no unemployment in last 10 years, time since last unemployment in last 10 years, mean employment duration in last 4/10 years, number of unemployment/out of labour force/programme spells in last 10 years, distance to hypothetical entry into labour market (calculated from age and education)
6	Benefits and UI claim	Amount of benefit, remaining UI claim, no UI claim
7	Compliance with benefit conditions, employability and mobility	Fully mobile within Germany, average job referrals per day, no referrals, at least one benefit sanction in past, at least one other type of non-compliance with benefit conditions in past
8	Health	Has health impairments, impairments affect employability, recognised disability status, total duration reported in sick during receipt of benefits in past, did not report in sick during receipt of benefits in past
9	Characteristics of job looked for	Skill profile, full/part-time, occupation
10	Regional information	GDP growth 1994-2002, travel time to next big city on public transport, fraction of foreigners, unemployment rate, agglomeration area, rural area, net migration

3.4 Relation to the data used in comparable studies

We claim that the German linked administrative employer-employee data we use is the most comprehensive dataset currently available for the evaluation of typical job search assistance and training programmes for the unemployed. Clearly, administrative data outperform any survey data available in terms of reliability, sample size, period covered, and representativeness. Moreover, compared to the survey data used in LaLonde (1986) and Dehejia and Wahba (1999) the set of available characteristics is considerably larger. Moreover, there are no comparable datasets that are suitable for the evaluation of active labour market

programmes which include detailed firm characteristics and allow constructing industry and occupation-specific work experience.[8]

In the following, we discuss a number of comparable studies based on quite informative administrative data that use selection-on-observable strategies to identify programme effects. With the exception of the linked firm information which have become available in Germany only very recently, administrative data in Germany are very similar to those available in Switzerland (see Gerfin and Lechner, 2002) and Austria (see Lechner and Wiehler, 2010). However, the data used in these studies are less informative with respect information regarding health and job search (characteristics of job looked for, vacancy referrals, compliance with benefit conditions). Yet, the Austrian data allow observing times in which females are on maternity leave, while we would only be able to classify the person as out of the labour force without being able to distinguish why. On the other hand, the Swiss data include a variable that provides a subjective caseworker assessment of the employability of each jobseeker, while we can capture this only indirectly with the number of vacancy referrals and the variable indicating whether there are health problems that affect employability. Similar information exist in the Swedish data used by Sianesi (2004) which contain the caseworker's assessment of the client's job readiness, need for guidance and difficulty to be placed. Yet, her data lack information on health, marital status, number and age of kids, occupation and skill profile of last job, firm characteristics of last job other than industry, occupation looked for and, importantly, on employment histories.

Another comparable study is Mueser, Troske and Gorislavsky (2007) who assess the performance of the U.S. JTPA programme using administrative data from Missouri. In con-

[8] Some datasets include the industry of the last job (e.g. Sianesi, 2004) and firm size (e.g. Lechner, Miquel and Wunsch, 2010). So far, linked employer-employee data is mainly used for other labour market analysis than the evaluation of labour market programmes (see Abowd and Kramarz, 1999).

trast to our data they are unable to control for health, marital status, number and age of kids, skill profile and industry of last job as well as other firm characteristics, anything related to job search, detailed regional variables as well as amount of benefits and UI claims. Moreover, they only observe employment histories up to two years before the intervention. Jespersen, Munch and Skipper (2008) use Danish administrative data to assess Danish labour market programmes. Although the data are in many ways similar to ours they lack information on health, occupation and skill profile of last job, firm characteristics, and anything related to job search.

The final set of related studies is comprised of studies using earlier versions of the German administrative data. The first generation of data, which covered training programmes, were used by Lechner, Miquel and Wunsch (2007, 2010) as well as Fitzenberger and Speckesser (2007) and Fitzenberger and Völter (2007). These data lack information on health, anything related to job search, and firm characteristics other than industry and firm size. The next generation of data is used, for example, in Lechner and Wunsch (2008, 2009). The data are the predecessor of the current version and cover a shorter period but are identical to the data we use here except that they lack the firm characteristics other than industry.

In summary, our data comprise the union of the information available in other comparable studies except for information on maternity leave in the Austrian data, and a case-worker assessment of the jobseeker in the Swiss and Swedish data. However, as argued above and in Section 3.3, we are able to capture the main aspects of this indirectly. Moreover, our data are even more informative and hence unique because they contain several measures of individual health and a variety of important firm characteristics. Finally, as can be seen from the list of variables in the Internet Appendix, we put considerable effort in capturing all as-

pects of individual employment histories by constructing a large variety of different measures from the data.[9]

3.5 Estimation

Since it has been argued above that identification of the average programme effects has been achieved by controlling for (almost) all potentially relevant confounding factors, a matching estimator is a natural choice as it allows for effect heterogeneity and does not require any functional for the relationship of the outcome variables and the selection variables (see for example the excellent survey by Imbens, 2004, and Imbens and Wooldridge, 2009). It is the common strategy in the literature on programme evaluation to tackle the dimensionality problem by instead of conditioning on the selection variables directly, to condition on an estimate of the conditional participation probability instead (the so-called propensity score, see Rosenbaum and Rubin, 1983). That estimate is typically performed using a parametric model, so that the full estimation procedure comes semiparametric. Here, we use a binary probit model for the propensity score. The full specification and the coefficient estimates for two propensity score models (using the full specification) are provided in the internet appendix. These models have been tested extensively against misspecification (non-normality, heteroscedasticity, omitted variables).[10]

The matching procedure used in this paper incorporates the improvements suggested by Lechner, Miquel, and Wunsch (2010). These improvements tackle two issues: (i) To allow for higher precision when many 'good' comparison observations are available, they incorporate the idea of calliper or radius matching (e.g. Dehejia and Wahba, 2002) into the standard

[9] Of course, not all of them are included in the selection models but we extensively test for omitted variables.

[10] The test results as well as the results for further specifications that are used in the following sections are available on request from the authors.

algorithm used for example by Gerfin and Lechner (2002). (ii) Furthermore, matching quality is increased by exploiting the fact that appropriately weighted regressions that use the sampling weights from matching have the so-called double robustness property. This property implies that the estimator remains consistent if either the matching step is based on a correctly specified selection model, or the regression model is correctly specified (e.g. Rubin, 1979; Joffe, Ten Have, Feldman, and Kimmel, 2004). Moreover, this procedure should reduce small sample as well as asymptotic bias of matching estimators (see Abadie and Imbens, 2006) and thus increase robustness of the estimator. The exact structure of this estimator is shown in Table B.1 in Appendix B.

There is an issue here on how to draw inference. Abadie and Imbens (2008) show that the 'standard' matching estimator (nearest neighbour or fixed number of comparisons) is not smooth enough and, therefore, bootstrap-based inference is not valid. However, the matching-type estimator implemented here is by construction smoother than the one studied by Abadie and Imbens (2008) because we have a variable number of comparisons and because we apply the bias adjustment procedure on top. Therefore, it is presumed that the bootstrap is valid. It is implemented following MacKinnon (2006) by bootstrapping the p-values of the t-statistic directly based on symmetric rejection regions.[11]

Two issues affecting the appropriateness of matching estimators are common support with respect to the propensity score, and match quality. If there is insufficient common support in the different states, no appropriate matches are at hand for a subset of observations. For this reason, we discard any observation in one state having a higher or lower propensity score estimate than, respectively, the maximum or minimum in the other state. This, of

[11] Bootstrapping the p-values directly as compared to bootstrapping the distribution of the effects or the standard errors has advantages because the 't-statistics' on which the p-values are based may be asymptotically pivotal whereas the standard errors or the coefficient estimates are certainly not.

course, affects the population the causal effects refer to given that discarded observations systematically differ from the original sample. If the sample size is considerably reduced due to the common support restriction, one might therefore argue that the effects are not representative for the target population any more. Fortunately, due to large and heterogeneous pool of non-participants, common support is not an issue in this study. In fact, only one participant in a job search programme and two participants in a training programme have been removed. To speed up the estimation a bit and base it on a more homogenous sample we also removed 4% of the comparison group to the job search assistance programme and 2.5% of the comparison group for the training programme, because those observations would never appear in any match. After this step, the propensity has been re-estimated on the common support.[12]

The match quality concerns the question whether the distribution of the confounders is balanced among matched observations in states implying that comparable individuals with respect to the confounder values were actually matched. Checking the means and medians of potential confounders for matched individuals in different states suggests that the after-match balance is high for both comparisons.

4. The benchmark specification

This specification includes all variables mentioned above. We will analyse the role of these variables for the propensity score as well as for the outcomes.

4.1 Selection into the programmes

In Table 4.1 we show the sample means of selected variables for participants and non-participant in each programme (see the Internet Appendix for a full list). We also display their

[12] There was no need to reiterate this procedure as no support problem appeared with re-estimated propensity score.

standardized difference in % in order to assess the magnitude of potential selection bias as proposed by Imbens and Wooldridge (2009). The displayed numbers are calculated for the final estimation sample, which has been restricted to the common support of the propensity scores of participants and non-participants that is estimated from the benchmark specification. A lack of overlap is defined in terms of propensity scores below (above) the minimum (maximum) score of the comparison group. Only 4% of the job-search-assistance sample and 3% of the training sample have been deleted in this step, and these were mainly elderly people, foreigners and jobseekers from regions with low unemployment rates.

The main insights from the standardized differences can be summarized as follows: Extreme selection as defined by Imbens and Wooldridge (2009) in terms of standardized differences above 25% exists only in very rare cases. Overall, as hinted at in Section 2, selection is generally stronger for training than for job search assistance: For the latter only 4% of the variables show a standardized difference above 20% and only 12% of the variables exceed 10%, while for training the respective fractions are 3% and 21%. For both programmes, selection is strongest in terms of age, region, unemployment duration at programme start, previous unemployment experience, whether there have been any vacancy referrals in the spell, health, and marital status. For job search assistance differences are also large for previous programme participation, out-of-labour-force experience in the long-term employment history, and - which is related to the latter - the difference between potential and actual work experience. In contrast, for training we find large differences for the variables indicating potential qualification needs, namely education, industry, skill profile and occupation of last job, as well as the occupation looked for.

Table 4.1: Descriptive statistics of selected variables for the different subpopulations

	Job search assistance			Training		
	Programme participants	Non-partici-pants	Absolute standardized difference in %	Programme participants	Non-partici-pants	Absolute standardized difference in %
Baseline characteristics						
Age	33.75	36.80	21	36.01	37.22	8
Female	0.39	0.40	2	0.47	0.40	10
No school degree	0.10	0.11	2	0.06	0.11	11
Lower secondary school degree	0.52	0.55	5	0.47	0.56	13
Upper secondary school degree	0.24	0.21	5	0.27	0.20	11
University entry degree	0.14	0.13	2	0.19	0.13	14
Foreign citizen	0.13	0.15	4	0.10	0.15	10
Married	0.37	0.45	12	0.43	0.46	4
Baden-Wurttemberg	0.12	0.13	2	0.14	0.13	1
Bavaria	0.10	0.21	22	0.16	0.22	11
Lower Saxony, Bremen	0.16	0.16	1	0.17	0.16	2
North-Rhine-Westphalia	0.27	0.27	1	0.26	0.26	1
Schleswig-Holstein, Hamburg	0.19	0.08	25	0.11	0.08	9
Hesse	0.07	0.08	2	0.08	0.08	1
Rhineland- Palatinate, Saarland	0.08	0.08	1	0.09	0.07	4
Timing of entry into unemployment and programme						
Beginning of unemployment	35.77	32.18	15	29.20	31.49	10
Time to treatment	6.82	5.17	21	7.92	6.27	22
Last employment: non-firm characteristics						
Earnings from last month per fortnight in EUR	742.00	755.69	2	811.40	758.44	8
Last job: unskilled	0.34	0.32	4	0.24	0.32	13
Last job: skilled	0.19	0.24	8	0.16	0.24	14
Last employment: firm characteristics						
Last job: firm size	255.03	295.61	2	250.38	295.38	2
Last job: fraction of jobs destroyed in firm	0.26	0.24	4	0.26	0.24	4
Short-term employment history (up to 2 years before unemployment)						
Fortnights employed in last year	15.61	15.22	3	16.39	15.09	11
Fortnights unemployed in last year	4.06	4.46	4	3.35	4.58	14
Long-term employment history (up to 10 years before unemployment)						
Fortnights employed in last 10 years	123.11	134.23	12	135.75	135.12	1
Fortnights unemployed in last 10 years	30.88	31.55	1	26.55	31.98	10
Fortnights out of labour force in last 10 years	83.84	72.15	12	75.55	70.82	5
At least one programme in last 10 years	0.27	0.21	9	0.26	0.21	7
Benefits and UI claim						
Benefit per fortnight in EUR	276.72	270.55	3	270.36	271.43	1
Remaining UI claim in days	286.25	314.63	10	320.48	319.25	0
Compliance with benefit conditions, employability and mobility						
No vacancy referral	0.16	0.33	28	0.20	0.33	21
Some benefit sanction in past	0.07	0.05	5	0.05	0.06	2
Some other form of noncompliance in past	0.15	0.12	5	0.10	0.12	5

Note: Table 4.1 to be continued.

Table 4.1 continued

	Job search assistance			Training		
	Programme participants	Non-participants	Absolute standardized difference in %	Programme participants	Non-participants	Absolute standardized difference in %
Compliance with benefit conditions, employability and mobility						
No vacancy referral	0.16	0.33	28	0.20	0.33	21
Some benefit sanction in past	0.07	0.05	5	0.05	0.06	2
Some other form of noncompliance in past	0.15	0.12	5	0.10	0.12	5
Health						
Has health impairments	0.16	0.21	9	0.15	0.21	12
Health impairments affect employability	0.08	0.12	9	0.07	0.12	12
Characteristics of job looked for						
Desired job: technical occupation	0.14	0.12	4	0.16	0.12	9
Desired job: construction occupation	0.14	0.16	4	0.08	0.16	17
Desired job: higher skilled service occupation	0.34	0.34	1	0.48	0.34	21
Desired job: lower skilled service occupation	0.17	0.17	0	0.12	0.17	9
Desired job: other occupation	0.20	0.19	1	0.14	0.19	10
Regional information						
Local GDP growth 1994-2002	19.80	20.77	6	20.64	20.85	1
Local unemployment rate	8.66	8.27	9	8.35	8.28	2

Note: Entries in the first two columns for each programme are sample means and represent fractions of observations if not stated otherwise. The standardized difference is defined as the difference in sample means of respective participants and corresponding non-participants divided by the square root of the sum of the empirical variances in the two subsamples. It is given in %.

4.2 Which variables do really matter?

In the last section we have shown for both programmes that participants and non-participants actually differ significantly in a number of characteristics. However, in order to identify programme effects we only need to control for those factors that have a joint impact on both selection into the programme and the outcomes of interest. In Tables 4.2 and 4.3 we therefore provide p-values for Wald tests of the joint significance of the 11 blocks of variables in the propensity-score estimation and the outcome equations for, respectively, job search assistance and training. For the outcome equations we estimate simple parametric models in the population of non-participants. It is important to note that their character is just illustrative to assess the relevance of the blocks of variables rather than an attempt to estimate the correct model. They are only used in these tables. As outcome variables we use different measures of employment status and earnings four years after (simulated) programme start.

Tables 4.2 and 4.3 clearly indicate that all blocks of variables we consider are strongly related to both selection into the programmes and all outcome variables. The only exception is the information on benefits and UI claim, which are jointly insignificant in the selection model for training. It is important to note that the tests indicate the relevance of a given block of variables conditional on all other blocks being included in the model. For benefit information and UI claims in the training equation, controlling for past earnings and the short and long-term employment history is likely to already capture some determinants of UI status. However, this also means that the high p-values indicate strong relevance for each individual block even given all the other blocks.

The alternative specifications of the selection models we are going to estimate, which are explained in detail in Section 5, all use subsets of the variables included in the benchmark model. In the lower part of Tables 4.2 and 4.3 we provide the results of Wald tests for the joint significance of the variables left out in one of those specifications, given the included once. They are based on the estimated benchmark model. Again, all blocks of excluded variables are highly significant, implying that leaving them out is likely to bias evaluation results and hence policy conclusions.

Table 4.2: Wald tests for blocks of variables in outcome equation and propensity score: Job search assistance (p-values in %)

Outcome equations	4 years after programme start		Average in year 4 after programme start		Cumulated effects over the first 48 months after programme start				Propensity score
Blocks of variables	employment rate in %	monthly earnings	months employed in %	monthly earnings	months employed	earnings in EUR	months unemployed	benefit receipt from UI	
Timing of entry into UE & progr. (1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Non-firm info. of last employment (2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3
Firm info. of last employment (3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info. on last employment (o_{lastem}, 2, 3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info. on last short employ. (4)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Long term employment history (5)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Employment history (o_{ehist})	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
History information (o_{hist})	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Current UI benefit information (6)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.6
Job search, mobility, employability (7)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Health impairments (8)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5
Desired job information (9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
Detailed regional information (10)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info on search activities (o_{search} 7,9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info on current UE (o_{ue}, 1, 6)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Standard information ($_{stan}$ 3,7,8,9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Variables not in Lalonde (1986)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dehejia & Wahba (1999)									
Sianesi (2001)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heinrich et al. (2009)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline demographics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: The outcome equations are based on linear regressions with the exception of the outcome *employment after year 4*, which is based on a probit. All test statistics related to the outcome equations are based on the subsample of non-participants.

Table 4.3: Wald tests for blocks of variables in outcome equation and propensity score: Training (p-values in %)

Outcome equations	4 years after programme start		Average in year 4 after programme start		Cumulated effects over the first 48 months after programme start				Pro-pensity score
Specifications	em-ploy-ment rate in %	monthly earn-ings	months em-ployed in %	monthly earn-ings	months em-ployed	ear-nings in EUR	months unem-ployed	benefit receipt from UI	
Timing of entry into UE & progr. (1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Non-firm info. of last employment (2)	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Firm info. of last employment (3)	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info. on last employment (o_{lastem}, 2, 3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info. on last short employ. (4)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Long term employment history (5)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Employment history (o_{ehist})	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
History information (o_{hist})	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Current UI benefit information (6)	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	85
Job search, mobility, employability (7)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Health impairments (8)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
Desired job information (9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Detailed regional information (10)	2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	5
Info on search activities (o_{search} 7,9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Info on current UE (o_{ue}, 1, 6)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Standard information ($_{stan}$ 3,7,8,9)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Variables not in Lalonde (1986)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dehejia & Wahba (1999)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sianesi (2001)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Heinrich et al. (2009)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline demographics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: See note below Table 3.1.

4.3 The effects of the different programmes

Table 4.4 displays the estimated effects of participating in job search assistance and training on various labour market outcomes for their respective participants. The first two outcomes in columns 2 and 3 measure whether a person is employed and the corresponding earnings (zero if not employed) in the last period (fortnight) of our observation window four years after (simulated) programme start. The employment outcome is a measure of longer-term integration into the labour market, while the earnings effect is a crude measure for changes in productivity. For job search assistance it could be interpreted as changes in job match quality while for training as changes in human capital. Columns 4 and 5 show

smoothed versions of these outcomes that are calculated as averages of these variables over all periods in the fourth year after programme start.

The last four columns of Table 4.4 display cumulated outcomes over the full four-year period after programme start. Usually, labour market programmes for the unemployed tend to have negative (positive) effects on employment and earnings (unemployment and benefit receipt) in the short run because of reduced job search activity during programme participation (so-called lock-in effects; see van Ours, 2004, Lechner, Miquel and Wunsch, 2010). The cumulated outcomes can be used as proxy for some kind of net effect of the programmes as they would trade off potential initial negative short-run effect against potential positive long-run effects (or vice versa).

Table 4.4: Estimated effects for job search assistance and training

Outcome variables	4 years after programme start		Average in year 4 after programme start		Cumulated effects over the first 48 months after programme start			
Specification of propensity score	employment rate in %	monthly earnings	months employed	monthly earnings	months employed	earnings in EUR	months unemployed	benefit receipt from UI
				Job search assistance				
Average effect for participants	0.1	-61	-0.9	-64	-5.6	-5130	-1.4	-554
p-value of two sided test	95	8.9	46	1.5	0.0	0.0	0.0	0.0
				Training				
Average effect for participants	3.0	55	2.5	48	-2.7	-919	-1.9	-577
p-value of two sided test	0.5	2.0	0.0	5.0	0.0	36	0.0	0.0

Note: The p-values are derived from the bootstrap distribution of the t-statistic assuming a two-sided and symmetric distribution. 199 bootstrap replications. The respective t-values are computed assuming matching weights as fixed and ignoring the regression step in the matching algorithm. The matching algorithm and the computation of the standard errors is explained in Lechner, Miquel, Wunsch (2010).

The results in Table 4.4 show no employment effects and weak evidence for small negative effects on earnings for job search assistance four years after the programmes. This is in line with other findings in the literature (see Martin and Grubb, 2001, Kluve, 2006). If job search assistance is effective, the effects are usually short-term and there is no evidence so far that it increases match quality. Moreover, summarized over the full four-year period employment and earnings of participants are significantly lower than those of comparable

non-participants. However, participants accumulate also fewer months of unemployment and benefit payments. This hints at an increase of other forms of non-employment.[13] For training we find small positive effects on employment and earnings after four years, but the cumulated effects are still negative, pointing to the presence of lock-in effects of these programmes. However, we again see a reduction in cumulated unemployment and benefits.

5. Sensitivity of the results to the particular specification

5.1 Does the available selection information matter for the policy conclusions?

The previous section revealed that almost all groups of variables discussed above should not be excluded from the model based on the test statistics of the t- and Wald tests. However, for a practical application these statistics do not reveal the full story, because they relate to the (ideal) situation in which only one of the blocs of variables is missing. In many applications such a scenario will be too optimistic. Furthermore, although leaving out one of the blocs may lead to biased estimates, the bias may be so small that it does not change the policy conclusion of the respective evaluation study. Therefore, in this section, we will systematically vary the available blocs of confounding variables and investigate whether the conclusion for the evaluation study would change as measured by the effects of the programmes on the various outcome variables defined in the previous section.

The first set of scenarios corresponds to the case considered in the previous section, namely that one bloc of variables is left out and all others are kept. The second set corresponds to just the opposite, namely the case when one particular bloc of variables is included on top of a set of baseline variables that could be expected to be available in every typical

[13] This has been found also in other studies, see e.g. Lechner, Miquel and Wunsch (2010).

study. Of course, the cases without any covariates or with just the set of baseline variables are considered as well. In addition to these specifications we consider some combinations of variable groups that can be expected to be frequently found in applications. These combinations correspond to the situation when (i) no employment history is available, when (ii) no information on the last job (firm and non-firm) is available, when (iii) no history information is available (no information on the last job and no employment history), when (iv) no information about job search is available, and when (v) no information about the current unemployment spell (timing information, benefits, UI claim) is available. Correspondingly we also look at the cases when this particular information is the only one available in addition to the base line covariates. Another specification is the one that only includes a standard set of variables found in many studies that contains standard individual characteristics, regional information, information on the last job (non-firm) and on individual employment histories but excludes non-standard variables related to firm information, health, and job search.

Finally, we look at combination of different variable groups to mimic the variables available for a couple of papers that have been or might become influential. The first one is the seminal paper by LaLonde (1986) who only used a minimum set of control variables to assess the performance of matching estimators (age, gender, education, race, earning in the two years before programme start). We also look at the adjusted version proposed by Dehejia and Wahba (1999) who add marital status and more pre-treatment information on employment status. The next study is the one by Sianesi (2004). She uses rich Swedish administrative data but compared to our benchmark specification she lacks information on health, marital status, number and age of kids, occupation and skill profile of last job, firm characteristics of last job other than industry, and the occupation the unemployed is looking for. Moreover, she has no information on employment histories. However, she uses a sample of individuals who have never been unemployed before. Consequently, she controls for unemployment experi-

ence but cannot capture total, industry and occupation-specific work experience, job stability and times out of the labour force. Yet, in contrast to our data, Sianesi (2004) observes variables that indicate the caseworker's assessment of the client's job readiness, need for guidance and difficulty to be placed, which we can only control for indirectly, as described in Section 3.3. We also look at the study by Mueser, Troske and Gorislavsky (2007) who assess the performance of matching and matching combined with a difference-in-difference approach relative to experimental estimates of the effect of the U.S. JTPA programme. They use rich administrative data from Missouri but are unable to control for health, marital status, number and age of kids, skill profile and industry of last job as well as other firm characteristics, anything related to job search, detailed regional variables as well as amount of benefits and UI claims. Moreover, they only observe employment histories up to two years before the intervention. The final study considered is Lechner, Miquel and Wunsch (2010) where we use the first administrative data set available in Germany for the evaluation on training programmes. These data lack information on health, anything related to job search, and firm characteristics other than industry and firm size.

Tables 5.1 and 5.2 show the results for the average programme effects for the programme participants corresponding to the different scenarios. First, consider the two extreme cases, namely the full model and the unadjusted raw difference. Considering the earnings and employment outcomes in year 4, for JSA it appears that due to the adjustment for covariates the raw differences of an 8%-point employment gain and 12 EUR monthly earnings gain become zero for employment and significantly negative for earnings once all confounding variables are controlled for. This adjustment clearly suggests that JSA participants are in general positively selected. We observe a similar phenomenon for training for which the large raw differences in the employment rate and earnings shrink by more than half once all confounders are controlled for. Indeed, the fact that the programme effects get much worse when all

confounders are controlled for, can be observed for both programmes and all outcome variables.[14]

The next step is now to consider the impact of the different variable *combinations.*[15] Here, the surprising insight is that controlling for the baseline specification is removing already a big part of the bias due to confounding. This is also probably the reason why it is hard to detect effect differences across the specifications that are statistically significant.

[14] As 'worse' we define lower earnings and employment and higher unemployment and benefit receipt.

[15] Note that from the statistical point of view there is no reason to expect that by adding new confounders one should smoothly come closer to the estimate that occurs when all confounders are controlled for. The magnitude and direction of the change due to additional information depends on the correlation structure of the included and excluded confounders and the outcomes.

Table 5.1: Estimated effects for job search assistance

Outcome variables	4 years after programme start		Average in year 4 after programme start		Cumulated effects over the first 48 months after programme start			
Specification of propensity score	employment rate in %	monthly earnings	months employed	monthly earnings	months employed	earnings in EUR	months unemployed	benefit receipt from UI
All variables included	0.1	-61	-0.9	-64+	-5.6+	-5130+	-1.4+	-554+
- no timing of entry into UE and progr. (1)	3.1+	-13	1.4	-27	-3.3+	-3667+	-1.7+	-670+
- no non-firm info. of last employment (2)	-0.5	-53	-1.7	-60+	-5.4+	-4328+	-1.3+	-492+
- no firm info. of last employment (3)	0.0	-39	-0.0	-52+	-4.7+	-4574+	-1.5+	-605+
- no info. on last employ. (o_lastem, no 2, 3)	0.3	-48	-0.9	-47+	-5.4+	-4430+	-1.4+	-551+
- no long term employment history (5)	0.0	-61+	-1.3	-67+	-5.4+	-4695+	-1.4+	-520+
- no employment history (o_ehist)	1.7	-11	0.3	-18	-4.1+	-3301+	-1.8+	-613+
- no history information (o_hist)	1.9	-15	0.3	-29	-4.4+	-3655+	-1.9+	-767+
- no current UI benefit information (6)	-0.3	-58	-1.2	-62+	-5.8+	-4719+	-1.4+	-484+
- no job search, mobility, employability (7)	1.3	-46+	-0.7	-58+	-5.0+	-4566+	-1.3+	-463+
- no health impairments (8)	0.6	-50	-0.9	-55+	-5.1+	-4398+	-1.2+	-418+
- no desired job information (9)	0.4	-42	-0.7	-45+	-4.9+	-4178+	-1.5+	-554+
- no detailed regional information (10)	1.2	-44+	-0.6	-59+	-4.9+	-4650+	-1.4+	-535+
- no info on search activities (o_search no 7,9)	0.7	-47	-0.7	-61+	-5.1+	-4887+	-1.0+	-406+
- no info on current UE (o_ue, no 1, 6)	1.7	-23	0.3	-37	-4.4+	-3833+	-1.7+	-493+
Only *standard* information (stan no 3,7,8,9)	2.3	-17	0.9	-27	-4.1+	-3832+	-1.2+	-471+
Specification similar to Lalonde (1986)	1.9+	-51+	-0.6	*-79+*	**-6.5+**	**-6832+**	-1.7+	-658+
Dehejia and Wahba (1999)								
Sianesi (2004)	1.5	3	0.7	-10	-3.9+	-2757+	-1.5+	-527+
Mueser, Troske and Gorislavsky (2007)	0.5	-23	-0.6	-39+	-5.2+	-4063+	-1.2+	-367+
Lechner, Miquel and Wunsch (2010)								
Only baseline demographics	1.7	-23	-0.5	-51+	-5.9+	-5032+	-1.9+	-624+
- & timing of entry into UE and progr. (1)								
- & non-firm info. of last employment (2)								
- & firm info. of last employment (3)								
- & info. on last employ. (o_lastem, no 2, 3)								
- & long term employment history (5)								
- & employment history (o_ehist)								
- & history information (o_hist)								
- & current UI benefit information (6)								
- & job search, mobility, employability (7)								
- & health impairments (8)								
- & desired job information (9)								
- & detailed regional information (10)								
- & info on search activities (o_search no 7,9)								
- & info on current UE (o_ue, no 1, 6)								
Raw difference on common support	4	12	2	30	-3.9	-4578	-4.1	-1556

Note: +: Effect is significantly different from zero at 5% level (based on bootstrapping t-values). Bold: Difference with specification *all variables included* is significantly different from zero at the 5% level (based on bootstrap distribution of difference). Italics: Difference with specification *all variables included* is significantly different from zero at the 10% level (based on bootstrap distribution of difference). 199 bootstrap replications. All estimates are based on the same common support.

Note for SOLE conference referee: As computation is time intensive, we have not yet completed all simulations. The remaining parts will be added before the conference starts. Inference for the final estimates will also be based on 499 bootstrap replications in-

stead of the 199 used now. In this sense our current conclusions are preliminary and may be subject to change.

Table 5.2: Estimated effects for training

Outcome variables	4 years after programme start		Average in year 4 after pro-gramme start		Cumulated effects over the first 48 months after programme start			
Specification of propensity score	employ-ment rate in %	monthly earnings	months employ-ed in %	monthly earnings	months em-ployed	earnings in EUR	months unem-ployed	benefit receipt from UI
All variables included	3.0+	55+	2.5+	48	-2.7+	-919	-1.9+	-577+
- no timing of entry into UE and progr. (1)	3.6+	60+	2.3+	43	-2.7+	-1087	-1.7+	-429+
- no non-firm info. of last employment (2)	2.8+	55+	2.2+	53+	-2.9+	-514	-1.9+	-533+
- no firm info. of last employment (3)	2.8+	46+	1.8	43+	-2.9+	-1093	-1.8+	-533+
- no info. on last employ. (o_{lastem}, no 2, 3)	3.3+	50	2.2+	39	-2.8+	-1121	-1.7+	-468+
- no long term employment history (5)	4.6+	66+	3.4+	56+	-2.1+	-596	-2.1+	-649+
- no employment history (o_{ehist})	3.6+	43	2.6+	48+	-2.7+	-1152	-2.4+	-798+
- no history information (o_{hist})	3.9+	70+	2.8+	54+	-2.7+	-848	-2.3+	-687+
- no current UI benefit information (6)	2.4	39	1.9+	36	-3.2+	-1501	-1.7+	-477+
- no job search, mobility, employability (7)	3.3+	37	2.6+	33	-2.8+	-1488	-1.6+	-437+
- no health impairments (8)	2.5	40	2.1+	37	-2.8+	-1392	-1.7+	-505+
- no desired job information (9)	2.7+	39	2.0+	39	-2.6+	-723	-1.7+	-476+
- no detailed regional information (10)	1.9	14	1.4	13	-3.4+	-2465	-1.7+	-531+
- no info on search activities (o_{search} no 7,9)	2.9+	28	1.9	23	-2.9+	-1494	-1.6+	-446+
- no info on current UE (o_{ue}, no 1, 6)	2.6	58+	1.9	43+	-2.8	-937	-1.9+	-544+
Only *standard* information ($stan$ no 3,7,8,9)	3.8+	52+	2.6+	44	-2.5+	-935	-1.8+	-538+
Specification similar to Lalonde (1986) Dehejia and Wahba (1999) Sianesi (2004) Mueser, Troske and Gorislavsky (2007) Lechner, Miquel and Wunsch (2010)	4.4+	86+	2.6+	59+	-2.0+	-110	-2.0+	-557+
Only baseline demographics	4.5+	78+	2.6+	58+	-3.5+	-1059	-1.7+	-452+
- & timing of entry into UE and progr. (1)								
- & non-firm info. of last employment (2)								
- & firm info. of last employment (3)								
- & info. on last employ. (o_{lastem}, no 2, 3)								
- & long term employment history (5)								
- & employment history (o_{ehist})								
- & history information (o_{hist})								
- & current UI benefit information (6)								
- & job search, mobility, employability (7)								
- & health impairments (8)								
- & desired job information (9)								
- & detailed regional information (10)								
- & info on search activities (o_{search} no 7,9)								
- & info on current UE (o_{ue}, no 1, 6)								
Raw difference on common support	8	160	7	131	0	2201	-4.2	-1431

Note: See note below Table 5.1.

5.2 The reason of the differences

In the previous section we analysed the difference in the effect estimates that come with the different specifications. Here, we want to analyse the reasons for the difference in

some more detail. Clearly, biases come about when the omitted variable bias of the propensity score also affects the outcome variable. Table 5.3 considers the first issue by showing how much the propensity score that is obtained from models with fewer confounders is correlated with the propensity score of the full model. If that correlation is 100%, there won't be any bias. If the correlation is zero, then we expect to get no adjustments and obtain the raw difference instead. As an additional measure of predictive power we also compare the Pseudo-R^2's that are obtained in the different specifications, because a loss of prediction of the true propensity score matters only if it is related to a loss of predictive power for the outcome equation of the respective non-participants. Therefore, Table 5.4 presents the R^2 and Pseudo-R^2 that correspond to the particular outcome variables for JSA.[16] Of course, these equations may be misspecified because of the implied functional form assumptions of this parametric model and thus not fully informative about the changes in effect estimates observed in the previous section. Nevertheless, the variation in the R^2, but not necessarily its absolute level, should still be informative about the relevance of the particular blocs of variables.

From Table 5.3 we see immediately that we do not expect much difference in the effect estimates by *separately* omitting variable groups (2), (3), (4), (3 & 4), (5), (6), (8), (9), (10) as the other variables will probably pick up their correlation concerning the selection into the programmes. Similarly, with one exception, Table 5.4 shoes that the reduction in the R^2 for these specifications compared to the full specification is small as well. Given the richness of the specification, this is of course not surprising. The exception to the latter finding is the

[16] The results presented are for the non-participants of job search assistance. However, since the non-participation samples for both programmes are very similar, the effects for training-non-participants look almost identical and are referred to the internet appendix. Further note that we report R^2 not *adjusted* R^2 because we are interested the predictive value of the omitted group independent of how variables this takes. Anyway, due to the large sample size the difference between unadjusted and unadjusted R^2's are very small.

amount of benefit received which is (of course) well predicted by the variables denoting the remaining benefit claim (6).

Further interpretation when the final results have been obtained.

Table 5.3: Analysis of selection model

Programmes	Job search assistance		Training	
Specification of propensity score	Correlation of p-score with base in %	Pseuo-R^2 of p-score in %	Correlation of p-score with base in %	Pseuo-R^2 of p-score in %
All variables included	-	8.1	-	5.9
- no timing of entry into UE and progr. (1)	81	5.1	87	4.4
- no non-firm info. of last employment (2)	100	8.0	98	5.7
- no firm info. of last employment (3)	98	7.9	98	5.6
- no info. on last employ. (o_lastem, no 3,4)	98	7.9	95	5.4
- no long term employment history (5)	98	7.8	97	5.6
- no employment history (o_ehist)	94	7.3	93	4.9
- no history information (o_hist)	92	6.9	85	4.3
- no current UI benefit information (6)	100	8.1	100	5.9
- no job search, mobility, employability (7)	94	7.0	95	5.4
- no health impairments (8)	100	8.1	100	5.8
- no desired job information (9)	100	8.0	99	5.7
- no detailed regional information (10)	99	7.8	100	5.8
- no info on search activities (o_search no 7,9)	93	7.0	94	5.2
- no info on current UE (o_ue, no 1, 6)	79	4.8	87	4.4
only *standard* information ($stan$ no 3,7,8,9)	92	6.7	90	4.8
Specification similar to Lalonde (1986)	30	0.6		
Dehejia & Wahba (1999)				
Sianesi (2001)	95	7.4	89	4.7
Heinrich et al. (2009)	65	3.4		
Only baseline demographics	53	2.2	47	1.3
- & timing of entry into UE and progr. (1)				
- & non-firm info. of last employment (2)				
- & firm info. of last employment (3)				
- & info. on last employ. (o_lastem, no 2, 3)				
- & long term employment history (5)				
- & employment history (o_ehist)				
- & history information (o_hist)				
- & current UI benefit information (6)				
- & job search, mobility, employability (7)				
- & health impairments (8)				
- & desired job information (9)				
- & detailed regional information (10)				
- & info on search activities (o_search no 7,9)				
- & info on current UE (o_ue, no 1, 6)				

Note: Efron's R^2 is used as Pseudo-R^2. It is based on a comparison of the maxima of the likelihood function in a model with a constant term compared to a model with all regressors.

Note for submitted paper: Some specifications are still missing (because of the large sample, use of the bootstrap and a semiparametric estimator, computation time is large) and will be added before the conference.

Table 5.4: R^2 in the outcome equations of the non-participants for job search assistance

Outcome variables	4 years after programme start		Average in year 4 after programme start		Cumulated effects over the first 48 months after programme start			
Specification of control variables	employ-ment rate in %	monthly earnings	months employ-ed	monthly earnings	months em-ployed	earnings in EUR	months unem-ployed	benefit receipt from UI
All variables included	13	20	17	25	26	33	41	51
- no timing of entry into UE and progr. (1)	12	20	16	24	26	33	40	49
- no non-firm info. of last employment (2)	12	19	17	23	26	31	41	51
- no firm info. of last employment (3)	12	20	16	24	26	32	41	51
- no info. on last employ. (o_lastem, no 2, 3)	12	19	16	23	25	30	41	51
- no long term employment history (5)	12	19	16	24	25	32	40	50
- no employment history (o_ehist)	11	18	14	23	22	30	39	49
- no history information (o_hist)	10	17	13	20	21	26	38	48
- no current UI benefit information (6)	12	20	16	25	26	33	37	39
- no job search, mobility, employability (7)	12	20	17	25	26	33	41	51
- no health impairments (8)	12	20	16	24	25	33	41	51
- no desired job information (9)	12	19	16	24	26	32	41	51
- no detailed regional information (10)	12	20	17	25	26	33	41	51
- no info on search activities (o_search no 7,9)	12	19	16	24	26	32	40	51
- no info on current UE (o_ue, no 1, 6)	12	19	16	24	25	32	36	38
Only *standard* information ($stan$ no 3,7,8,9)	12	19	16	24	25	31	40	51
Specification similar to Lalonde (1986)	6	9	7	10	10	12	19	20
Dehejia & Wahba (1999)								
Sianesi (2001)	10	18	13	22	21	29	40	50
Heinrich et al. (2009)	10	17	13	21	21	28	32	35
Only baseline demographics	7	11	9	13	12	16	20	22
- & timing of entry into UE and progr. (1)								
- & non-firm info. of last employment (2)								
- & firm info. of last employment (3)								
- & info. on last employ. (o_lastem, no 2, 3)								
- & long term employment history (5)								
- & employment history (o_ehist)								
- & history information (o_hist)								
- & current UI benefit information (6)								
- & job search, mobility, employability (7)								
- & health impairments (8)								
- & desired job information (9)								
- & detailed regional information (10)								
- & info on search activities (o_search no 7,9)								
- & info on current UE (o_ue, no 1, 6)								

Note: The outcome equations are based on linear regressions with the exception of the outcome *employment after year 4*, which is based on a probit. All outcome equations are estimated in the subsample of non-participants. Efron's R^2 is reported for the probit, all other R^2 do not adjust for the sample size and number of covariates.

5.3 Heterogeneity

As the literature on the evaluation of labour market programmes suggest that there are substantial difference between man and women (see e.g. the survey by Bergemann and van den Berg, 2006), in this section we investigate whether the observed differences of the effects

hold for men and for women in the same way, or whether the specific groups of variables affect the programme effects for men and women differently.

Note for submitted paper: Gender specific results not yet complete. They will be added before the conference.

6. Conclusion

In this paper we analyzed the role different confounding variables play for the estimation of the effects of typical programmes of active labour market policies by matching methods. As our benchmark we use the probably most comprehensive data base that is currently available for these types of evaluation studies. It is a large and newly enriched German administrative data base. When understanding on how the different variables influence the results, we analyzed the effect on the propensity score, which is a necessary but not sufficient condition for being a relevant determinant of the final effect estimate, as well as their impact on the outcomes directly.

Our preliminary results suggest that the role of the different variables vary significantly and may be quite different compared to what has been reported in the literature so far (based on different, far less informative data sets). However, at this stage the results are still too incomplete to come to definitive conclusions.

….

Apparently, there are also several caveats with our approach. First, although the German programmes we look at, namely job search assistance and vocational training, are typical programmes for many countries, there are always some differences from one country to the other with respect to the working of the programme and the selection process. However, they should be small at least for the big European countries and may be taken into account if these

results for Germany are generalized to other settings. Second, one can never proof that indeed all factors are covered. However, given the richness of the data, it is difficult to image factors missing that could to further substantial biases.

References

Abadie, Alberto, and Guido W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267.

Abadie, Alberto, and Guido W. Imbens (2008): "On the Failure of the Bootstrap for Matching Estimators", *Econometrica*, 76, 1537-1557.

Abowd, J. and F. Kramarz (1999): "The Analysis of Labor Markets using Matched Employer-Employee Data," Handbook of Labor Economics, O. Ashenfelter and D. Card (eds.), Chapter 26, Vol. 3B, North-Holland, 2629-2710.

Bergemann, Annette, Gerard van den Berg (2006): "Active Labor Market Policy Effects for Women in Europe: A Survey", Discussion Paper 2365, Institute for the Study of Labor (IZA).

Blundell, Richard, and Monica Costa Dias. 2009. "Alternative Approaches to Evaluation in Empirical Microeconomics." *Journal of Human Resources* 44(3): 565–640.

Card, David, Jochen Kluve, and Andrea Weber (2009): "Active Labor Market Policy Evaluations: A Meta-Analysis," IZA DP 4002.

Dehejia, Rajeev (2005): "Practical propensity score estimation: a reply to Smith and Todd", *Journal of Econometrics*, 125, 355-364.

Dehejia, Rajeev H., and Sadek Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes", *Journal of the American Statistical Association, 94, 1053-1062.*

Dehejia, Rajeev H., and Sadek Wahba (2002): "Propensity score-matching methods for nonexperimental causal studies", *Review of Economics and Statistics*, 84, 151-161.

Dorsett, Richard (2006): "The new deal for young people: effect on the labour market status of young men", *Labour Economics*, 13 , 405–422.

Fredriksson, Peter, and Per Johansson, (2003): "Program Evaluation and Random Program Starts", Discussion Paper 2003(1), IFAU, Uppsala.

Fredriksson, Peter, and Per Johansson, (2008): "Dynamic Treatment Assignment - The Consequences for Evaluations Using Observational Studies", *Journal of Business Economics and Statistics* 26(4): 435-445.

Gerfin, Michael, and Michael Lechner (2002): "A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", *The Economic Journal*, 112, 854-893.

Heckman, James J., and V. J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, 862-880.

Heckman, J.J. and J.A. Smith (1995): "Assessing the Case for Social Experiments", Journal of Economic Perspectives, 9, 85-110.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program", *Review of Economic Studies*, 64, 605-654.

Heckman, James J., Robert LaLonde, and Jeffrey A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, 3, 1865-2097, Amsterdam: North-Holland.

Heckman, James J., and Jeffrey A. Smith (1999): "The pre-programme dip and the determinants of participation in a social programme: Implications for simple programme evaluation strategies", *Economic Journal* 109: 313–348.

Heckman, James J. , and Jeffrey A. Smith (1999): "The Determinants of Participation in a Social Program: Evidence from a Prototypical Job Training Program", *Journal of Labor Economics*, 22, 243-298.

Imbens, Guido W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review", *The Review of Economics and Statistics*, 86, 4-29.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5-86.

Jespersen, Svend T., Jakob Roland Munch, and Lars Skipper (2008): "Costs and Benefits of Danish Active Labour Market Programmes", *Labour Economics*, 15, 859-884.

Kluve, J. (2006): "The Effectiveness of European Active Labor Market Policy", IZA Discussion Paper 2018, Institute for the Study of Labor (IZA), Bonn.

LaLonde, R.J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review*, 76, 604-620.

Lechner, M., R. Miquel, and C. Wunsch (2010): "Long-Run Effects of Public Sector Sponsored Training in West Germany", *Journal of the European Economic Association*, forthcoming.

Lechner, Michael, and Stephan Wiehler (2010): "Kids or Courses? Gender Differences in the Effects of Active Labor Market Policies", *Journal of Population Economics*, forthcoming.

Larsson, Laura (2003): "Evaluation of Swedish Youth Labor Market Programs." *Journal of Human Resources* 38(4).

MacKinnon, J. G. (2006): "Bootstrap Methods in Econometrics", *The Economic Record*, 82, 2-18.

Mueser, P. R., K. Troske and A. Gorislavsky (2007): "Using State Administrative Data to Measure Program Performance", *Review of Economics and Statistics*, Vol. 89, 761-83.

OECD (2008): OECD Employment Outlook 2008. Paris.

Peikes, Deborah N., Lorenzo Moreno, and Sean Michael Orzol (2008): "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs," *The American Statistician*, 62 (3), 222-231.

Petrongolo, Barbara (2009): The long-term effects of job search requirements: Evidence from the UK JSA reform, *Journal of Public Economics* 93 (2009) 1234-1253.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

Shadish, William R., M. H. Clark, and Peter M. Steiner (2008): "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments, *Journal of the American Statistical Association, Applications and Case Studies*, 103 (484), 1334-1356.

Sianesi, Barbara (2004), "An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s", *Review of Economics and Statistics*, 86, 133-155.

Smith, Jeffrey A., and Petra Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?", *Journal of Econometrics*, 125, 305-353.

Thomsen, Stephan (2009): "Job Search Assistance Programs in Europe: Evaluation Methods and Recent Empirical Findings", FEMM Working Paper No. 18, Otto-von-Guericke University Magdeburg.

Van Ours, Jan (2004): "The Locking-in Effect of Subsidized Jobs", *Journal of Comparative Economics*, 32, 37-52.

Appendix A: Data

Will be added.

Appendix B: Technical details of the matching estimator used

Table B.1: A matching protocol for the estimation of a counterfactual outcome and the effects

Step 1	Specify a reference distribution defined by X.
Step 2	Pool the observations forming the reference distribution and the participants in the respective period. Code an indicator variable *D*, which is 1 if the observation belongs to the reference distribution. All indices, 0 or 1, used below relate to the actual or potential values of *D*.
Step 3	Specify and estimate a binary probit for $p(x) := P(D=1 \mid X = x)$
Step 4	Restrict sample to common support: Delete all observations with probabilities larger than the smallest maximum and smaller than the largest minimum of all subsamples defined by *D*.
Step 4	*Estimate the respective (counterfactual) expectations of the outcome variables.* **Standard propensity score matching step (multiple treatments)** a-1) Choose one observation in the subsample defined by *D*=1 and delete it from that pool. b-1) Find an observation in the subsample defined by *D*=0 that is as close as possible to the one chosen in step a-1) in terms of $p(x), \tilde{x}$. 'Closeness' is based on the Mahalanobis distance. Do not remove that observation, so that it can be used again. c-1) Repeat a-1) and b-1) until no observation with *D*=1 is left. **Exploit thick support of X to increase efficiency (radius matching step)** d-1) Compute the maximum distance (*d*) obtained for any comparison between a member of the reference distribution and matched comparison observations. a-2) Repeat a-1). b-2) Repeat b-1). If possible, find other observations in the subsample of *D*=0 that are at least as close as $R * d$ to the one chosen in step a-2) (to gain efficiency). Do not remove these observations, so that they can be used again. Compute weights for all chosen comparisons observations that are proportional to their distance. Normalise the weights such that they add to one. c-2) Repeat a-2) and b-2) until no participant in *D*=1 is left. d-2) For any potential comparison observation, add the weights obtained in a-2) and b-2). **Exploit double robustness properties to adjust small mismatches by regression** e) Using the weights $w(x_i)$ obtained in d-2), run a weighted linear regression of the outcome variable on the variables used to define the distance (and an intercept). f-1) Predict the potential outcome $y^0(x_i)$ of every observation using the coefficients of this regression: $\hat{y}^0(x_i)$. f-2) Estimate the bias of the matching estimator for $E(Y^0 \mid D=1)$ as: $\sum_{i=1}^{N} \frac{1(D=1)\hat{y}^0(x_i)}{N^1} - \frac{1(D=0)w_i\hat{y}^0(x_0)}{N^0}$. g) Using the weights obtained by weighted matching in d-2), compute a weighted mean of the outcome variables in *D=0*. Subtract the bias from this estimate to get $\widehat{E(Y^0 \mid D=1)}$.
Step 5	Repeat Steps 2 to 4 with the nonparticipants playing the role of participants before. This gives the desired estimate of the counterfactual nonparticipation outcome.
Step 6	The difference of the potential outcomes is the desired estimate of the effect with respect to the reference distribution specified in Step 1.

The parameter used to define the radius for the distance-weighted radius matching (*R*) is set to 90%. This value refers to the distance of the worst match in a one-to-one matching and is defined in terms of the propensity score. Different values for *R* are checked in the sensitivity analysis in Lechner, Miquel, and Wunsch (2010). The results were robust as long as *R* did not become 'too large'.