CESifo Working Papers

www.cesifo.org/wp

Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables

Michael Lechner
Conny Wunsch

CESIFO WORKING PAPER NO. 3381
CATEGORY 4: LABOUR MARKETS
MARCH 2011

An electronic version of the paper may be downloaded
- *from the SSRN website: www.SSRN.com*
- *from the RePEc website: www.RePEc.org*
- *from the CESifo website: www.CESifo-group.org/wp*

CESifo *Working Paper No. 3381*

Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables

Abstract

Based on new, exceptionally informative and large German linked employer-employee administrative data, we investigate the question whether the omission of important control variables in matching estimation leads to biased impact estimates of typical active labour market programs for the unemployed. Such biases would lead to false policy conclusions about the cost-effectiveness of these expensive policies. Using newly developed Empirical Monte Carlo Study methods, we find that besides standard personal characteristics, information on individual health and firm characteristics of the last employer are particularly important for selection correction. Moreover, it is important to account for past performance on the labour market in a very detailed and flexible way. Information on job search behaviour, timing of unemployment and program start, as well as detailed regional characteristics are also relevant.

JEL-Code: J680.

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

Michael Lechner
Swiss Institute for Empirical Economic Research / University of St. Gallen
St. Gallen / Switzerland
Michael.Lechner@unisg.ch

Conny Wunsch
Swiss Institute for Empirical Economic Research / University of St. Gallen
St. Gallen / Switzerland
Conny.Wunsch@unisg.ch

This draft: March 2011
Date this version has been printed: 10 March 2011
Michael Lechner is also affiliated with CEPR and PSI, London, CES-Ifo, Munich, IAB, Nuremberg, IZA, Bonn, and ZEW, Mannheim. Conny Wunsch is also affiliated with CES-Ifo, Munich, and IZA, Bonn. This project received financial support from the Institut für Arbeitsmarkt und Berufsforschung, IAB, Nuremberg (contract 8104), and from the St. Gallen Research Center for Aging, Welfare, and Labor Market Analysis (SCALA). Parts of the paper were written while the first author visited CES-Ifo, Munich, and the second author visited UCL, London. The hospitality and support of both institutions is gratefully acknowledged. We would like to thank Patrycja Scioch (IAB), Benjamin Schünemann and Darjusch Tafreschi (both SEW, St. Gallen) for their help in the early stages of data preparation. The paper has been presented at seminars in Frankfurt, St. Gallen, and CEMMAP, London, as well as at the EALE/SOLE meeting in London, and the second meeting of the Danish Microeconometric Network in Skagen. The usual disclaimer applies.

1. Introduction

Costing up to 3% of GDP (OECD, 2010) active labour market policies aiming at bringing the unemployed back to work belong to the most important public expenditure programs in OECD countries. Thus, there is considerable and increasing interest among both policy makers and researchers to quantify the effects of these programs on the labour market outcomes of their participants.

Since, for good reasons, the participants in specific active labour market programs are not a random selection of jobseekers, all empirical studies attempting to evaluate the effect of such programs face the problem of so-called selection bias. In the absence of social experiments, an increasing number of evaluation studies argue that the data they use are informative enough to remove selection bias by controlling for observed variables.[1] The key assumption in these studies is that the data contain all variables that jointly influence outcomes, typically post-program earnings and employment indicators, and program participation. If this assumption is true, controlling for these 'confounding' variables will identify particular average effects of these programs with a minimum number of further assumptions required. Generally, the econometric methods used in this literature are well advanced. Many benchmark applications exist.[2]

Many governments have become aware of the value of informative and accurate data to obtain reliable impact estimates that form the basis for any subsequent cost-benefit analysis. Hence, they are making their administrative databases available to the scientific community that uses them extensively. Although the rich content of these data, which varies somewhat from one country to

---

[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, Mueser, Troske, Jeon, Kahvecioglu (2009) for the USA.

[2] The methodological advances are summarized in the comprehensive surveys by Blundell and Costa-Dias (2009) and Imbens and Wooldridge (2009). Card, Kluve, and Weber (2009), for example, cover the large recent applied literature in their meta-study.

the next, is the main justification for the validity of the empirical results obtained by econometric matching methods, it is surprising that there is not yet any systematic investigation of exactly which variables are required to avoid substantial selection biases in such studies.

A convincing investigation of this issue requires knowledge, or a credible benchmark estimate, of the effect of interest, as well as observing in the data all factors that may cause a spurious correlation between program participation and the outcomes of interest. Moreover, to be of broader relevance such an investigation should focus on typical selection problems (i.e. typical programs using typical assignment rules). So far, most of the existing literature uses social experiments conducted in the U.S.[3] to obtain a benchmark estimate of the effect of interest (LaLonde, 1986, Fraker and Maynard, 1987, Friedlander and Robins, 1995, Heckman and Smith, 1999, Dehejia and Wahba, 1999, 2002; Heckman, Ichimura, and Todd, 1997, 1998, Heckman, Ichimura, Smith, and Todd, 1998, and Smith and Todd 2005).[4] These studies match the experimental data to another non-experimental dataset and then compare the result using the experimental control group with the results using the non-experimental control group.

Social experiments might indeed be a good reference case if the implementation is unproblematic, if they have a large enough sample to determine the 'truth' precisely, and if they are representative for the programs of interest. However, most social experiments do not meet all of these requirements (Heckman and Smith, 1995). Also, a much more serious problem is that existing experimental datasets do not contain the necessary wealth of information argued to be required for

---

[3] No experimental benchmarks exist for European programs. The main reasons behind this lack of evidence are ethic concerns when denying jobseekers services that are deemed to be helpful in a random and thus arbitrary fashion. In the absence of experimental evidence, an alternative benchmark could be obtained from so-called quasi-experimental studies, for example obtained by instrumental variable estimation (e.g., Frölich and Lechner, 2010) or difference-in-difference estimation (e.g., Petrongolo, 2009). However, when program effects are heterogeneous, which is likely for active labour market programs, these methods identify instrument specific parameters. As different studies use different instruments, any cross-study comparison is difficult.

[4] See also Heckman and Hotz (1989), Peikes, Moreno, and Orzol (2008), and Shadish, Clark, and Steiner (2008).

credible selection correction in applications. Such information is at least partially included, though, in (most) of the datasets used for observational studies based on matching methods (see footnote 1). Thus, an important caveat of studies with an experimental benchmark is that they are unable to mimic realistic assignment decisions simply because the required variables are missing in their matched non-experimental data. Therefore, it is only possible to investigate how well the rather few observed covariates are balanced and how close the non-experimental estimates come to the experimental benchmark. There is no way to judge the importance of specific information on particular parts of the assignment process on the final impact estimates. Hence, this exercise is uninformative for real world selection problems occurring with active labour market programs because of the missing link to a realistic assignment process.

The objective of this paper is to improve on the important methodological dimensions discussed above and to provide a systematic investigation of the question which groups of variables are most likely required as control variables for classical evaluation studies of typical active labour market programs. We argue that the new German administrative linked employer-employee database we use contains information on all major factors claimed to be important for selection correction and that were used in the various applications that rely on the selection-on-observables assumption. We base our analysis on a design that is similar to the concept of an Empirical Monte Carlo Study proposed and applied in Huber, Lechner, and Wunsch (2010). The chosen design ensures that the true effect is known by construction rather than by assumption. This is a clear advantage, compared to unreliable or imprecise benchmark estimates. The design further ensures that we know the true selection model and that the unconfoundedness assumption holds in the (partially simulated) data. Moreover, the basis of our true selection model is an estimate of the selection probability of a typical application, which makes it much more realistic. We impose no assumptions about the relation between covariates and outcomes but exploit their dependencies in the data. We also argue that the programs we analyse for West Germany, namely job search assistance and training, are not only the most widely used programs in OECD countries but are also typical in terms of their contents, implementation, and selection of participants. Finally, our data contains the outcome variables typically used in evaluation studies.

We find that the availability of information on the health of the unemployed workers and on the firm characteristics of their last employer is particularly important for justifying the selection-on-observables assumption. Ignoring this information leads to considerable bias. Moreover, controlling for the timing of unemployment and program start as well as information on the job search behaviour is important as well. We also confirm the findings from the earlier literature that underlines the relevance of accounting for caseworker assessments (Gerfin and Lechner, 2002, Sianesi, 2004), pre-treatment outcomes (Mueser, Troske, and Gorislavsky, 2007), transitions between different labour market states and detailed regional information (Friedlander and Robins, 1995, Heckman, Ichimura, Smith, and Todd, 1998, Heckman and Smith, 1999), as well as for labour market histories in a flexible way (Dolton and Smith, 2010). We also argue that the lack of important control variables is likely to impact on cost-benefit analyses and may therefore lead to wrong policy conclusions regarding the cost-effectiveness of the programs.

The remainder of the paper is organized as follows: The next section describes in detail the programs we analyze, how they compare to other countries and how selection of participants works. Section 3 outlines the research design. In Section 4, we provide all details on the data used, their relation to other datasets available, and argue that they justify the identification of program effects by a selection-on-observables assumption. We also describe the matching methods used. Section 5 analyses the selection into the programs and describes our empirical selection model. Section 6 presents the results. The last section concludes. An Appendix as well as an additional Internet Appendix (contained in the online discussion paper version of this paper) contains further details on the data and the estimation.

2. The determinants of participation in typical labour market programs

2.1 Programs 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 programs 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* programs implemented in Germany is very representative for this class of programs (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. In the latter, jobseekers learn how to locate job vacancies, how to write an application and practice job interviews.

German *training* programs are heterogeneous. They include those types of programs commonly used in most other OECD countries,[5] but other programs differ with respect to the form and intensity of the human capital investment involved and their respective duration (ranging from several weeks to more than two years). Therefore, we restrict our analysis to the subgroup of programs that are internationally most typical. They comprise occupational skills training, skill upgrading and programs combining workplace training with related instruction that have planned durations of no more than six months.

The implementation of the two types of German programs we look at is also largely representative with respect to eligibility and selection into the programs. Job search assistance is used

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

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 is targeted more specifically towards jobseekers with certain qualification deficits. In the period we consider, 2000-2002, eligibility for program participation required jobseekers to qualify for unemployment insurance (UI) payments (so-called unemployment benefits), or for unemployment assistance, 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 programs and their participants in the period we consider here (2000-2002).

2.2 Participation in the programs

In general, program participation is the outcome of decisions made by both the caseworker and the unemployed person. Usually the caseworker proposes participation in a program to improve a client's reemployment prospects, though sometimes the jobseeker also proposes a program. In either case, the jobseeker must apply for permission before beginning any subsidised program. The caseworker decides whether 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 what kind of program, if any, would be appropriate. An assessment of the jobseeker's employment prospects and the specific qualification needs is the basis for this decision. According to the German legislation, caseworkers also have to take into account the chances of a successful completion of the program, as well as the local labour market conditions. Similar arguments apply to the decision making process of the unemployed. They most likely compare their employment prospects with and without a specific program, as well as the corresponding costs in terms of required effort and alternative use of time. Their decision to accept the participation decision made by the caseworker should also consider potential sanctions in case of non-compliance.

Similar to many other countries, there are institutional incentives to participate in labour market programs. Jobseekers refusing to participate in a program they were assigned to risk a benefit sanction, i.e. a temporary cut or withdrawal of their unemployment benefit or unemployment assistance. Moreover, for our period of investigation, and this is a feature mainly of some European countries, participation in a *training* programs stops the clock for exhausting UI benefits, i.e. the remaining maximum UI benefit duration at the beginning of a program is the same as at the end.[6] Since there are also benefit payments during the program, jobseekers effectively extend their potential UI benefit duration by participating in a program. This feature, however, does not occur for job search assistance, where the unemployed, if eligible, continue to receive their UI benefit and potentially use up their UI claim.

These considerations on the selection process have the following implications for strategies identifying the effects of job search assistance and training programs on labour market outcomes by selection-on-observables assumptions: First, it is important to note that all determinants of program participation mentioned above are likely to affect labour market outcomes like employment status and earnings as well. Thus, they are potential confounders that have to be measured and used as control variables.

The first measurement issue is to ensure eligibility for program participation. To do so, we have to determine whether unemployed individuals qualify for unemployment benefits or assistance. Moreover, to capture institutional incentives we must observe the level of benefits, UI eligibility status, and the remaining potential UI benefit duration. 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, health, motiva-

[6] 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.

tion 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 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 jobseeker 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 program completion, essentially the same factors play a role as for employment prospects and qualification needs. The final set of factors relates to preferences and alternative ways of using the time out of employment. The most relevant cases are probably women's fertility decisions, the main determinants of which would have to be captured. In particular, Lechner and Wiehler (2010) show that program participation and becoming pregnant during unemployment are both attractive options for women. For men alternative time use may be less important because institutions provide strong incentives to leave unemployment, making the leisure value of unemployment less relevant.

3. Empirical design

When assessing the role of covariate information for matching-based program evaluations it is of key importance to have a credible benchmark against which different specifications of the selection correction model can be judged. For the analysis, we use the observed group of individuals who did not participate in any programs in a pre-specified period. In this group, we simulate a placebo treatment for which we know that by construction the true program effect is zero. The placebo treatment is assigned based on a selection model that is guided by actual selection decisions, i.e. it is estimated from actual participants and nonparticipants in the program (as would be done in an application). Next, based on the estimated participation model, the participation probability for a given program is predicted for each nonparticipant. A fraction of nonparticipants that is equal to the fraction of actual participants in the data is assigned randomly to the placebo treatment conditional on the predicted participation probability. This procedure ensures that the true selection model is known but as close as possible to real selection decisions,[7] and that the unconfoundedness assumption holds. It is worth pointing out that this simulation procedure imposes no assumptions about the relations between covariates and outcomes.[8]

To analyze the sensitivity of the estimated program effects with respect to the specification of the selection model, we re-estimate the effects including or leaving out different blocs or combinations of blocs of variables that are part of the true selection model. Repeating the simulation-estimation procedure 500 times for each specification allows us then to estimate the joint empirical distribution of the specification-specific estimators.[9]

To assess the role of sampling error and the implications for actual applications we also estimate all specifications without any simulation, i.e. we use the actual data of participants and nonparticipant relying on the validity of the unconfoundedness assumption using the specification of the 'true' model in the placebo data.

---

7 The coefficients of the model estimated from actual participants and nonparticipants become the true selection parameters in the placebo data.

8 Jacob, Ludwig, and Smith (2009) analyze random assignment of housing vouchers among applicants. They apply a similar approach. They use the sample of non-applicants and randomize out applicants for which the effect of the vouchers is known to be zero. They then redefine the treatment of interest as having applied for vouchers and study the performance of different matching estimators and different specifications of the selection model for estimating the effect of interest. The important difference to our approach is that the unconfoundedness assumption may be violated in their data, and that they are unable to consider the actual treatment of interest, namely receiving the voucher. Khwaja, Picone, Salm, and Trogdon (2010) apply an idea that is similar in spirit but more different in detail to a health intervention. They simulate under the assumption that the treatment effect is known using estimates of a structural model.

9 To ensure that the samples are independent, we first draw with replacement 500 samples of the same size as the original placebo data and then simulate participation within those 500 samples.

4. Data and econometric methodology

4.1 Data

We use a unique linked employer-employee administrative database. It is probably *the* most informative database that is currently available for evaluating typical labour market programs (see Section 4.5 for a discussion of how this data compares to other available data). Our data comprise a 2% random sample drawn from the population of all German employees subject to social insurance[10] since 1990. It covers the period 1990-2006 and combines information from different administrative sources: (1) the records provided to the social insurance system by employers for each employee (1990-2006), (2) the unemployment insurance records (1990-2006), (3) the program participation register of the Public Employment Service (PES, 2000-2006) as well as (4) the job-seeker register of the PES (2000-2006). Because these records determine social insurance and unemployment benefit claims as well as program eligibility, the data are very accurate with respect to employment status, earnings from employment, amount and duration of UI claims, and program participation status. The information collected by the PES on jobseekers is reliable 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 spells 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) the population of all employees of the firm as of June 30 of each year from 1990

[10] This covers 85% of the German workforce. It excludes the self-employed and civil servants.

to 2006 in which the firm existed (so-called establishment history panel or *Betriebshistorikpanel*, BHP). Finally, a variety of regional information was matched to the data via the official codes of the 439 German districts (*Kreiskennziffer*). It contains the population density, migration and commuting streams, average earnings, GDP growth, the unemployment rate, the share of long-term unemployment, welfare dependency rates, urbanisation, as well as childcare and public transport facilities.

For each individual the data comprise all aspects of their employment, earnings and UI history since 1990. This includes the first and last day of each spell, the type of employment (full/part-time, high/low-skilled), the occupation, earnings, the type and amount of UI benefit, and the remaining potential UI benefit duration. Furthermore, it includes the information about compliance with the benefit conditions (e.g. failure to show up at interview, refusal to participate in assigned labour market program, imposition of sanction), and periods when a UI recipient has reported being sick to the UI. Moreover, they cover all spells of participation in the major German labour market programs from 2000 onwards with exact beginning, end and type of program as well as the planned end date of the training programs. 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 he 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.

4.2 Sample selection and definition of participation status

Since we are interested in evaluating typical labour market programs in industrialized economies, we restrict the analysis to former West Germany (without Berlin). We start with a sample that covers all entries into unemployment in the period 2000-2002. Then, we exclude unemployment entries in January-March 2000. This is because we want to ensure that we do not accidentally classify entries from subsidized employment (in particular employment programs) as entries from unsubsidized employment due to a potential lack of an accompanying program spell.[11] Furthermore, we restrict the analysis to the population aged 20-59 in order to avoid having to model educational choices or (early) retirement decisions. We also ensure eligibility for program participation by requiring individuals to qualify for unemployment benefits or unemployment assistance. Finally, we exclude a few cases that start their unemployment spell directly with some program or for whom the information from the jobseeker register is missing.

As in Lechner, Miquel, and Wunsch (2010), we define as (non-) participants all those individuals in our sample who (do not) start a program within the first 12 months of their unemployment spell.[12] To focus on the internationally most widely used types of programs, we only consider participants whose first program is job search assistance, or training with a planned duration of no more than six months. This excludes atypical training programs that are unusually long compared to other countries.

In order to determine time to treatment and to measure outcomes relative to program start we simulate hypothetical program start dates for nonparticipants by drawing randomly from the empirical distribution of start dates of program 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

[11] The program information starts only in January 2000 and is not fully reliable in the first quarter of the year 2000.

[12] Nonparticipation means not starting any program in the 12-month window, not just the program used for the particular comparison.

distributions of start dates.[13] This implies that we have different samples of nonparticipants for job search assistance and training. We then impose hypothetical program eligibility on nonparticipants by requiring them to be unemployed and eligible for unemployment benefits or assistance at simulated program start. [14] Moreover, we discard all individuals with actual or hypothetical program start after 2002 to ensure that outcomes can be observed for up to four years after program start.

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

Although our research design guarantees the validity of the selection-on-observables assumption in the placebo data, to be relevant the selection model used should be plausible. Moreover, since we will also use the actual data on participants and nonparticipants to assess the implications for actual applications, plausibility of the unconfoundedness assumptions lends credibility to the conclusions drawn from this supplementary exercise.

At the end of Section 2, we summarized all factors that should be controlled for when identifying causal effects of the two programs on labour market outcomes based on a selection-on-observables approach. Here, we briefly relate them to the available data and discuss them in turn: *Eligibility* for program participation is ensured by the construction of the sample. Concerning the *institutional incentives*, we directly observe the amount of benefits, the UI eligibility and the remaining potential UI benefit duration. To measure *local labour market conditions* we observe the rich set of regional indicators listed in Section 4.1 that allow controlling for the relevant regional differences in a detailed way.

---

[13] Job search assistance is used very early in the spell while training starts later.

[14] Related to the arguments of Fredriksson and Johansson (2003, 2008), Sianesi (2004), and Lechner and Wiehler (2011) this definition of non-participation raises issues about dynamic program assignment and future labour market outcomes of the so-defined nonparticipants. However, as long as we condition on time to treatment, it does not affect our ability to model selection into the programs given the data. Moreover, we are only interested in comparing different models for selection correction and all our specifications will be based on the same treatment definition.

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 with the skill profile of the last job held. We approximate productivity by the earnings from the last job (controlling for full/part-time) and by the average earnings from employment in the last 10 years before current unemployment. In addition, we observe several variables indicating health problems, and variables indicating whether such problems affect employability. Work, occupation and industry-specific experience is 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, we use 10 years of detailed employment histories to control for the quality and stability of employment, for the frequency and duration of previous unemployment experience, and for other periods of non-employment. Second, we condition on the characteristics of the last employer that may reveal specific types of workers. Third, we control for incidence of non-compliance with benefit conditions during past unemployment spells. Fourth, we account 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. Finally, we know whether the jobseeker is fully mobile within Germany.

In addition to the factors like skills, productivity, experience, and motivation that were already mentioned, 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 to proxy for the *determinants of 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 women remain, of course, unobserved. However, we capture them indirectly to the extent to which they affect 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 aspects of fertility decisions and child raising preferences.

Table 4.1: Blocs of control variables

No.	Bloc	Variables
0	Baseline characteristics	Age, school degree, vocational degree, nationality, number of kids, age of youngest child <6, marital status
1	Timing of entry into unemployment & program	Half-month & quarter of entry into unemployment, time to treatment, interaction terms
2	Last employment: non-firm characteristics	Skill profile, full/part-time, occupation
3	Last employment: firm characteristics	Firm age, size, closed firm, fraction females, low-income, temporary & part-time jobs, age distribution, mean tenure, fraction of jobs destroyed, industry, most frequent occupation
4	Short-term labour market history (up to 2 years before unemployment)	Half-month employed/out of labour force (olf)/ in program in the 6/24 months before, no employment/unemployment in last 2 years, time since last unemployment/olf in last 2 years, unemployed/olf in month 6/24 before, number of unemployment/olf spells employer changes
5	Long-term labour market history (up to 10 years before unemployment)	Half-month employed/unemployed in the last 10 years before, in program/fortnights olf in the last 4/10 years before, no unemployment/olf in last 10 years, time since last unemployment/olf in last 10 years, mean employment/unemployment/olf duration in last 10 years, number of unemployment/out of labour force/program spells/employer changes in last 10 years, difference between potential & actual labour market experience, total time in last firm
6	Earnings history	Earnings in last job, average earnings in last 10 years, sum of earnings in last year/2 years
7	Industry & occupation-specific experience	Number of occupation/industry changes, tenure in last occupation/industry, total duration in last occupation/industry
8	Pre-treatment outcomes	Employed/earnings 4 years before, cumulated employment/earnings/ UI receipt/UI benefits over 4 years before
9	Benefits & UI claim	Amount of benefit, remaining potential UI benefit duration, no UI claim
10	Compliance with benefit conditions, employability & mobility	Fully mobile within Germany, average job referrals per day, no referrals, at least one type of non-compliance with benefit conditions in past
11	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
12	Characteristics of job looked for	Skill profile, full/part-time, occupation
13	Region dummies	State (*Bundesland*)
14	Detailed 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

In summary, perhaps with the exception of some aspects of preferences, our unique data enables us to capture the important confounding factors that affect both program participation and labour market outcomes. Thus, the selection-on-observables assumption appears to be credible.

Table 4.1 summarizes the blocs of variables that we use to control for selection. The choice of variables is driven by the identification arguments discussed above plus some specification tests (see Section 4.5). 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. Therefore, all estimations will be conducted separately for males and females (as well as for training and job search assistance).

4.4 Relation to the data used in comparable studies

We claim that the German administrative linked employer-employee data used here is the most comprehensive dataset currently available for the evaluation of typical job search assistance and training programs 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), Dehejia and Wahba (1999, 2002), Heckman and Smith (1999), Heckman, Ichimura, and Todd (1997, 1998), Heckman et al. (1998), and Smith and Todd (2005), the set of available characteristics is considerably larger. Moreover, there are no comparable datasets suitable for the evaluation of active labour market programs that include detailed firm characteristics and allow constructing industry and occupation-specific work profiles.[15] In the following, we discuss a number of studies based on quite informative administrative data that use selection-on-observable strategies to identify program effects.

[15] 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 used mainly for other labour market analysis than the evaluation of labour market programs (see Abowd and Kramarz, 1999).

With the exception of the linked firm information which have become available in Germany only 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 to 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 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) that 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 the last job, firm characteristics of the 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 JTPA program using administrative data from Missouri. In contrast 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 the amount of benefits and the 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 programs. Although their data is in many ways similar to our data, 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 programs, 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 (2009) and in Wunsch and Lechner (2008). The data are the predecessor of the version used here. They cover a shorter period but are identical to our data, except for the lack of 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 caseworker assessment of the jobseeker in the Swiss and Swedish data. However, as argued above and in Section 4.3, we 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 Appendix A, we put considerable effort in capturing all aspects of individual employment histories by constructing a large variety of different measures from the data.[16]

4.5 Estimation

Since we argued above that controlling for (almost) all potentially relevant confounding factors identifies average program effects, an econometric matching estimator is a natural choice. It allows for effect heterogeneity and does not require any specification of the functional relation of the outcome and the selection variables (see for example the excellent surveys by Imbens, 2004, and Imbens and Wooldridge, 2009). It is the common strategy in the literature on program evaluation to

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

tackle the dimensionality problem by conditioning on an estimate of the conditional participation probability (the so-called propensity score, see Rosenbaum and Rubin, 1983) rather than conditioning on the selection variables directly. This part of the estimation typically is performed using a parametric model, so that the full estimation procedure becomes semiparametric. Here, we use binary probit models for the propensity score. The full specification that uses all blocs in Table 4.1 and the coefficient estimates for all four propensity score models are provided in Appendix A. These models were tested extensively against misspecification (non-normality, heteroscedasticity, omitted variables).[17]

We use the matching estimator suggested by Lechner, Miquel, and Wunsch (2010) because it is one of the best estimators of a simulation study by Huber, Lechner, and Wunsch (2010). They compare the performance of all classes of propensity-score estimators typically used in practice: kernel matching, nearest and multiple-neighbour matching, inverse probability weighting, and parametric estimators. The estimator of Lechner, Miquel, and Wunsch (2010) incorporates the idea of calliper or radius matching (e.g. Dehejia and Wahba, 2002) into the standard algorithm used for example by Gerfin and Lechner (2002) to increase precision. Moreover, 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). The procedure reduces small sample bias as well as asymptotic bias of matching estimators (see Abadie and Imbens, 2006) and increases robustness. Appendix C describes the details of this estimator.

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

Two issues affecting the appropriateness of matching estimators are common support and match quality. If there is insufficient common support, then there is a subset of observations without appropriate matches. 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 course, affects the population the causal effects refer to given that discarded observations systematically differ from the original sample. If the sample size becomes considerably smaller due to the common support restriction, one might argue that the effects are not representative for the target population any more. Fortunately, due to a large and heterogeneous pool of non-participants, common support is not an issue here. In fact, only one participant in job search assistance and two training participants were removed. To speed up the estimation and to base it on a more homogenous sample we also removed 4% of the comparison group to the job search assistance program and 2.5% of the comparison group for the training program, because those observations would never appear in any match. After this step, the propensity score was re-estimated on the common support.[18]

Match quality relates to the question about the balance of the distribution of the confounders in the different treatment states. Checking means and medians of potential confounders for matched individuals in different states indicates that the after-match balance is high for all comparisons.

5. Selection into the programs

5.1 Descriptive statistics for the actual data

Table 5.1 presents sample means of selected variables for participants and non-participant in each program. We also display their absolute standardized difference in % in order to assess the

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

magnitude of potential selection bias as proposed by Imbens and Wooldridge (2009). The displayed numbers are calculated for the actual data, restricted to the common support.[19]

Table 5.1: Descriptive statistics of selected variables for the different subpopulations

		Job search assistance						Training					
		Men			Women			Men			Women		
		P	NP	SD	P	NP	SD	P	NP	SD	P	NP	SD
Age in years		33	37	24	34	38	24	35	37	10	37	38	7
Schooling:	No degree	.12	.12	0	.07	.08	3	.09	.11	5	.03	.08	14
	Upper secondary degree	.20	.15	9	.29	.28	1	.20	.16	8	.34	.28	8
	University entry degree	.12	.11	1	.17	.16	2	.18	.11	15	.21	.15	11
Vocational degree:	No degree	.37	.34	4	.32	.32	0	.27	.33	10	.22	.33	17
	University/college degree	.03	.04	2	.05	.05	1	.08	.03	14	.06	.05	5
Foreign citizen		.15	.17	3	.09	.11	6	.13	.15	5	.07	.12	12
At least one child		.24	.23	1	.38	.33	9	.23	.23	0	.41	.32	13
Married		.34	.43	12	.40	.49	13	.40	.43	4	.46	.48	3
Beginning of unemployment		37	32	18	35	32	12	30	32	7	28	31	13
Time to treatment in half-months		6.8	5.3	19	6.9	5.5	18	7.9	6.5	18	7.8	6.3	21
Remaining potential UI benefit duration in days		276	315	14	302	332	11	308	315	2	335	333	0
No vacancy referral		.16	.34	30	.17	.36	32	.18	.33	24	.22	.36	22
Any form of non-compliance with benefit conditions		.24	.19	9	.11	.10	3	.19	.19	0	.07	.10	7
Health problems (yes/no)		.17	.22	9	.15	.21	11	.16	.22	11	.14	.22	15
Looking for low- to medium-skilled job		.45	.43	3	.41	.40	2	.35	.42	10	.29	.41	18
Last job:	Half-monthly earnings in EUR	833	867	5	599	603	1	938	863	11	669	599	11
	Unskilled worker	.41	.37	5	.23	.21	3	.33	.37	6	.13	.22	16
	Clerk	.18	.16	4	.35	.35	0	.31	.16	27	.50	.35	21
	Firm size: # of employees	269	321	2	233	270	3	232	320	4	271	269	0
	Fraction laid off by firm	.27	.25	5	.24	.23	2	.26	.24	4	.26	.23	6
Cumulated over 2 years before:	# of UE spells	.65	.78	10	.43	.58	14	.61	.80	15	.39	.59	18
	# of out-of-labour-force spells	.80	.78	1	.72	.75	2	.68	.79	8	.63	.76	11
4 years before:	Employed	.56	.56	0	.51	.54	5	.58	.56	4	.57	.54	3
	Half-monthly earnings in EUR	786	910	9	564	627	6	920	900	1	669	625	4
Cumulated over 4 years before:	Employment	59	60	2	59	60	2	62	60	5	61	60	4
	Earnings in EUR/10000	52	57	9	38	40	4	61	57	7	44	40	9
	UI receipt	7.5	9.9	13	5.9	7.5	12	7.4	1.0	17	5.6	7.7	15
	UI benefits in EUR	1469	2038	16	809	1100	12	1430	2050	18	815	1122	13
Cumulated over 10 years before:	# of UE spells	1.7	2.1	14	1.0	1.3	14	1.5	2.1	19	.9	1.4	15
	# of out-of-labour-force spells	2.8	2.6	6	2.4	2.3	1	2.4	2.6	6	2.0	2.3	12
	# of occupation changes	3.7	3.3	10	2.9	2.7	6	3.6	3.3	8	2.7	2.7	3
	# of industry changes	2.2	1.9	13	2.0	1.8	8	2.1	1.9	10	1.8	1.8	1

Table 5.1 to be continued.

[19] The common support is obtained as explained in Section 4.5 using the propensity scores of participants and nonparticipants estimated from the actual data based on all blocs of control variables shown in Table 4.1.

Table 5.1 continued

	Job search assistance						Training					
	Men			Women			Men			Women		
	P	NP	SD	P	NP	SD	P	NP	SD	P	NP	SD
Baden-Wurttemberg	.12	.12	0	.13	.14	3	.12	.11	1	.15	.14	2
Bavaria	.09	.23	28	.12	.21	17	.15	.23	14	.17	.21	7
Lower Saxony, Bremen	.17	.16	2	.15	.15	0	.19	.16	5	.15	.15	1
Schleswig-Holstein, Hamburg	.19	.07	25	.20	.08	25	.11	.07	10	.11	.07	9
Hessen	.07	.08	2	.07	.08	3	.08	.08	0	.07	.08	2
Rhineland- Palatinate, Saarland	.08	.08	1	.07	.07	1	.08	.08	0	.11	.07	9
Local unemployment rate in %	8.8	8.3	12	8.5	8.2	7	8.5	8.3	5	8.2	8.2	0
# of observations	2267	32660		1452	22067		1754	30189		1570	20816	

Note: P: Mean among participants (fractions if not stated otherwise), NP: Mean among nonparticipants (fractions if not stated otherwise), SD: Absolute standardized difference in percent (difference in sample means of respective participants and corresponding nonparticipants divided by the square root of the sum of the empirical variances in the two subsamples).
Reference groups for dummies are omitted. 'before' and 'after' means before and after the beginning of the unemployment spell that determines membership in our population of interest. If not mentioned otherwise, all variables are measured at the beginning of this unemployment spell. Variables related to information in this spell are measured at the (simulated for controls) start of the program. Earnings are measured as earnings per half-month. 'Cumulated' measures sum up the half-monthly measures. Beginning of unemployment spell is measured in half-months where the first half of January 2000 equals '1'. All monetary measures are in EUR of the year 2000.

The variables displayed in Table 5.1 include the main baseline characteristics as well as the variables with the largest absolute standardized difference from each bloc of covariates. (see Tables A.1 in Appendix A for all variables included in the full selection model for both the actual and the placebo data).

The main insights from the standardized differences are 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 stronger for training than for job search assistance: For the latter 6-8% of all variables in Table 5.1 that will be included in the true selection model show a standardized difference above 15%, while for training the respective fraction is 10-11% (see Tables A.1 in Appendix A for the full table). For both programs, selection is strongest in terms of unemployment start, unemployment duration at program start, previous unemployment experience, vacancy referrals, health, and region. For job search assistance, differences are also large for age and marital status. In contrast, for training we find large differences for the variables indicating potential qualification needs, namely education, skill profile, and occupation of last job, as well as industry and the occupation looked for (for the last three variables see Tables A.1 in Appendix A).

5.2 Which variables do really matter?

For both programs, participants and nonparticipants differ significantly in a number of characteristics. However, in order to identify program effects we only need to control for those factors that have a joint impact on both selection into the program and the outcomes of interest. In Table 5.2, we therefore provide p-values for Wald tests of the joint significance of the 15 blocs of variables defined in Table 4.1 in the propensity-score estimation, and the outcome equations for both programs considered in the actual data. For the outcome equations, we estimate probit models for binary outcome variables and linear models for all other outcome variables in the population of nonparticipants.[20] It is important to note that the character of the outcome regressions is just descriptive to assess broadly the relevance of the blocs of variables. They are not an attempt to estimate the correct model and to derive causal conclusions; they are solely used for this table. As outcome variables, we use different measures of employment status and earnings four years after the (simulated) program start.

Table 5.2 indicates that all blocs of variables we consider are related strongly to selection into the programs and all outcome variables. There are only very few exceptions that mainly refer to women in job search assistance for whom program assignment seems to be less selective with respect to the characteristics of the last job, earnings history, UI eligibility and health. However, it is important to note that the tests indicate the relevance of a given bloc of variables conditional on all other blocs being included in the model. Thus, if we leave out one of the other blocs, these blocs can become important nevertheless. Therefore, we keep them in the analysis. Overall the low p-

[20] As this literature is usually interested in estimating the average effect of the program for the program participants, only the distribution of the characteristics of the non-participants has to be reweighted. Therefore, these regressions focus on non-participants only.

values indicate strong statistical relevance for each individual bloc even given all the other blocs, implying that leaving them out is likely to bias evaluation results and hence policy conclusions.

Table 5.2: P-values of Wald tests for the importance of blocs of variables

Blocs of variables	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Job search assistance - men														
Propensity score	0	0	2	0	0	0	1	14	5	2	0	6	3	0	0
Employed 4 years after	0	0	1	0	0	0	0	0	0	0	0	0	0	93	1
Half-monthly earnings 4 years after	0	0	1	11	0	0	0	0	11	0	0	0	0	8	0
Cumulated employment 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Cumulated earnings 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated UI receipt 4 years after in half-months	0	0	5	0	0	0	0	0	0	0	0	0	0	0	1
Cumulated UI benefits 4 years after	0	0	7	0	0	0	0	0	0	0	1	0	0	0	55
	Job search assistance - women														
Propensity score	0	0	37	15	2	0	21	5	5	16	0	64	3	0	0
Employed 4 years after	0	0	13	25	2	0	0	0	29	0	0	0	4	33	4
Half-monthly earnings 4 years after	0	0	12	15	7	0	0	0	0	0	2	0	0	0	0
Cumulated employment 4 years after in half-months	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated earnings 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated UI receipt 4 years after in half-months	0	0	3	0	0	0	0	0	0	0	0	0	0	0	2
Cumulated UI benefits 4 years after	0	0	4	0	0	0	0	0	0	0	0	0	2	1	39
	Training - men														
Propensity score	0	0	0	0	2	0	29	0	12	1	0	2	2	0	0
Employed 4 years after	0	0	0	0	0	0	2	0	0	0	0	0	0	56	4
Half-monthly earnings 4 years after	0	0	5	2	0	0	0	0	0	0	0	0	0	25	0
Cumulated employment 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated earnings 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated UI receipt 4 years after in half-months	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated UI benefits 4 years after	0	0	1	0	0	0	0	0	0	0	1	0	0	0	3
	Training - women														
Propensity score	0	0	0	0	0	1	7	18	7	1	0	12	0	0	0
Employed 4 years after	0	10	3	0	0	0	0	0	50	0	0	0	7	69	0
Half-monthly earnings 4 years after	0	0	24	8	3	0	0	0	0	0	1	0	0	0	0
Cumulated employment 4 years after in half-months	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated earnings 4 years after in half-months	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Cumulated UI receipt 4 years after in half-months	0	0	2	0	0	0	0	0	0	0	0	0	0	1	17
Cumulated UI benefits 4 years after	0	0	5	0	0	0	0	0	0	0	0	0	7	16	5

Note: Blocs 0 to 14 refer to the blocs of variables defined in Table 4.1.

5.3 The placebo data

As described in Section 3, the core of the analysis uses the subsample of actual nonparticipants in any program for whom the program effect is zero. For each group of the four groups (men/women in job search assistance/training), we use the actual data to estimate probit models for selection into the respective program using all variables in Table 4.1. Appendix A details those results. All blocs of variables shown in Table 4.1 are relevant for selection and the outcomes given all other blocs of variables at least in one of the four subsamples. Therefore, based on the specifications with all variables we predict the propensity score for each actual nonparticipant and randomly assign a placebo treatment based on this score such that the fraction of the simulated participants corresponds to the share of participants in the actual data. As explained above, this ensures that the model is realistic and relevant for applications. Moreover, the unconfoundedness assumption holds by construction. As expected, the means and standardized biases in the placebo data are similar to those of the actual data (see Internet Appendix I.1).

6. Results

The following subsections summarize the results from 57 specifications of the propensity score model in the four subsamples of men and women in training and job search assistance. Besides the full model, these specifications include 14 specifications where only one of the 14 blocs of variables is included besides the baseline characteristics (bloc 0 only), as well as 14 specifications where one of the 14 blocs of variables is excluded from the full model (all 15 blocs). In addition, we add to the model with the baseline characteristics and exclude from the full model groups of blocs of variables that comprise related factors like region dummies and detailed regional characteristics, firm and non-firm characteristics of the last job, and different combinations of labour market history variables. Finally, seven specifications mimic the specifications proposed in other studies. The tables provided in the Internet Appendix I.2 show the full list of specifications. We do not vary the variables within blocs because of computation time.[21]

[21] In total, we estimated 57 specifications, 500 times, in four subsamples, on both the simulated and the actual data, which adds up to 228'000 runs of the matching procedure.

We consider eight outcome variables that measure different dimensions. The majority of studies report employment rate and earnings at the end of the observation window (four years after program start in our case). We also report the averages of these variables over the last year yielding a smoothed version of the standard outcomes. The last set of outcomes provides a summary statistic for the whole observation period after program start: We cumulate the half-monthly outcomes over the full four-year period. We consider cumulated employment and earnings as well as cumulated unemployment and UI benefits. These outcomes provide some information on cost-effectiveness because they show the total returns in employment and earnings as well as potential cost savings in benefit payments and unemployment that can be contrasted with the direct program costs.

6.1 Importance of different blocs of variables

We use linear regressions to condense the information obtained from all specifications.[22] In Table 6.1, we specify the linear model such that a coefficient has the interpretation of the additional bias that occurs if a particular variable is removed from the full model (but all other blocs are retained), which by construction is unbiased. If a coefficient is positive, the estimated effect leaving out this bloc of variables is too large, i.e. it has an upward bias. Since the results are very similar across gender and program, we pool them.[23]

The results indicate that each bloc of variables significantly affects bias at least for some outcome variables. Information on health has the strongest single impact for all outcomes, followed by the characteristics of the last employer as well as earnings, unemployment and out-of-labour-force history, information of the timing of unemployment and program start as well as detailed re-

[22] These results are based on 50 specifications only, because the specifications that mimic other papers are left out because they do not correspond to a specific combination of the blocs of variables defined in Table 4.1.

[23] We also include a dummy for training and women in the pooled regressions. The gender and program-specific regressions are available on request.

gional characteristics. Moreover, information on individual job search effort, employability and mobility has a relatively large impact on the bias for the earnings outcomes.

Table 6.1: Regression results for the simulations

	4 years after program start		Average in year 4 after program start		Cumulated effects over the first 48 months after program start			
Bloc of variables removed from full specification	employ-ment rate in %	half-monthly earnings in EUR	months employed in%	half-monthly earnings in EUR	half-months employed	earnings in EUR	half-months on UI	benefit receipt from UI in EUR
Timing of entry into unemployment & program	*0.41*	*8.4*	*0.22*	*4.9*	*0.14*	*97*	*-0.24*	*-103*
Last job: Non-firm characterist.	*0.06*	*5.8*	0.02	*4.4*	0.01	*175*	*-0.06*	*-8*
Firm characteristics	*-0.11*	*-5.2*	*-0.31*	*-8.5*	*-0.44*	*-530*	*-0.06*	*-36*
Labour market history: 2 years	*0.15*	*-1.6*	0.02	*-4.1*	*-0.16*	*-348*	*-0.06*	*-28*
10 years	*-0.05*	*-3.7*	*-0.19*	*-6.6*	*-0.27*	*-425*	*-0.14*	*-72*
Earnings history	*0.26*	*10.5*	*0.28*	*11.3*	*0.31*	*581*	*0.11*	*69*
Industry- & occupation-specific experience	-0.06	*-4.3*	*-0.13*	*-5.6*	*-0.16*	*-297*	*-0.07*	*-37*
Pre-treatment outcomes	0.02	*2.0*	-0.05	0.6	*-0.13*	*-57*	*-0.03*	*-16*
Benefits & UI claim	*0.09*	*2.6*	*0.07*	*2.7*	-0.02	*89*	*0.07*	*51*
Compliance with benefit condit., employability & mobility	*0.04*	*-6.8*	-0.03	*-7.3*	*-0.17*	*-388*	*0.02*	*-6*
Health	*0.54*	*12.5*	*0.66*	*13.7*	*0.71*	*741*	*0.13*	*59*
Characteristics of job looked for	*0.08*	*3.0*	*0.04*	*1.6*	0.01	23	*-0.05*	*-23*
Region dummies	*-0.06*	*-6.1*	*-0.12*	*-7.7*	*-0.21*	*-439*	*-0.02*	*-35*
Detailed regional information	*-0.29*	*-5.7*	*-0.28*	*-5.6*	*-0.33*	*-285*	*-0.01*	*-4*
History: Employment	*0.08*	*2.0*	*0.09*	*2.2*	*0.07*	*101*	-0.02	-4
Unemployment	0.02	*-1.6*	*-0.13*	*-4.6*	*-0.35*	*-436*	*-0.08*	*-58*
Out-of-labour-force	*0.10*	*4.0*	*0.27*	*7.6*	*0.41*	*538*	*0.14*	*77*

Note: The entries refer to the mean - across simulations - of the coefficients of a regression of the bias (equal to the estimated effect because the true effect is zero) on dummies that is equal to one if the respective bloc of variables is left out in the estimation of the propensity score. *Italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level. Sample size for each regression: 200 observations (50 specifications x 4 subsamples). Standard errors obtained directly from the 500 simulation samples.
The first 14 blocs correspond to the blocs shown in Table 4.1. The last three blocs cover respective variables from the short-term and long-term labour market histories.

It is also interesting to consider related blocs of variables together. For the outcome 'employment rate in year 4', for example, we overestimate the program effects by half a percentage point if the health or the unemployment and program start information is missing (if both blocs were missing, the estimated program would be about one percentage point too large). Leaving out all of the regional information leads to an underestimate of the program effect by about a third of a percentage point, which pales compared to impact of the labour market history variables: Leaving them out leads to an overestimate of about two thirds of a percentage points. Finally, ignoring all information about the current unemployment spell biases the program effect by almost the same magnitude. Taken these results together, we conclude that every single bloc of variables is of limited impact, but when several blocs are missing, the biases may add up to substantial numbers. These findings are also confirmed by the more detailed results contained in Internet Appendix I.2, where we display the biases of the estimated effects of the programs in the placebo data for all specifications we consider. Often, these specifications leave out more than one bloc of variables. The results show again that biases generally increases the more information is omitted. Moreover, Wald tests based on the regressions presented in Table 6.1 reject specifications that leave out more than one bloc of variables (see again Internet Appendix I.3 for the p-values of these test statistics).

Table B.1 in Appendix B provides the corresponding regression results for the bias obtained from the estimation of the effects in the actual data. Here, bias is defined as the difference between the estimated effect from a given specification and the estimated effect of the full model that includes all variables in Table 4.1. There are two key differences to the simulations: The benchmark effect is not known but estimated and is therefore subject to sampling variation, and unconfoundedness does not hold by construction (but is plausible).

Considering the bias in this way allows us to relate it to the sampling error that would actually occur in an empirical study. Indeed, we find that sampling error in the benchmark estimate has a strong impact on the results. In contrast to the simulations, most coefficients are insignificant. However, the blocs of variables with the largest impact on bias in the simulations still appear as significant, at least for some outcomes: health, characteristics of last employer, timing information, unemployment, and out-of-labour-force history. Moreover, a closer look at the results reveals that sign and magnitude of the coefficients are very similar in the simulated and the actual data. Hence, it is unlikely that unobserved factors missing in the full model have a sizeable effect on both selection and the outcomes, because in this case their different correlations with the blocs of variables should lead to biases less consistent with the simulation results (for which the unconfoundedness assumption must hold).

6.2 Comparison with specifications used in other studies

In Table 6.2, we display the bias of the estimated effects of the programs in the placebo data for the true model as well as the specifications of the propensity score used in other studies for all subsamples.[24] We also report the correlation of the propensity score of the particular specification with the propensity score of the full model.[25]

We consider five benchmark specifications, all of which have considerably less information in several dimensions (see Section 4.4), but emphasize specific types of control variables: Sianesi (2004) underlines the importance of information about the caseworker's assessment of the job-seeker. Mueser, Troske, and Gorislavsky (2007) point to the importance of pre-treatment outcomes. LaLonde’s (1986) specification with the extensions proposed by Dehejia and Wahba (1999) is included as it is the standard benchmark in this literature despite having only a very limited set of control variables. Heckman and Smith (1999b) emphasize the importance of accounting for transitions between employment, unemployment and out-of-labour-force status as well as regional differences. Dolton and Smith (2010) advocate the necessity to control for labour market histories and transitions between labour market states in a flexible way. Since our full model controls for labour market histories in a very flexible way, we included two additional specifications where this information enters less flexibly.

---

[24] As the particular specifications were used to evaluate specific programs for specific groups of unemployed, it seems to give better justice to those specifications if subgroup specific results are displayed. Nevertheless, the differences between the four groups are small.

[25] Note that the bias for the true model is very close to zero implying that the chosen estimator performs very well. This finding is in line with the results obtained by Huber, Lechner, and Wunsch (2010).

Table 6.2: Bias of effects for selected specifications obtained from simulations

	Outcome variables	4 years after program start		Average in year 4 after program start		Cumulated effects over the first 48 months after program start			
Specification of propensity score	Correlation of *p(x)* with *p(x)* of full model	employ-ment rate in %	half-monthly earnings in EUR	months employ-ed in%	half-monthly earnings in EUR	half-months em-ployed	earnings in EUR	half-months on UI	benefit receipt from UI in EUR
		Training - men							
True model	1.00	0.0	-3	-0.1	-2	-0.06	-30	0.01	1
Sianesi (2004)	0.85	*1.4*	*41*	*1.1*	*38*	*1.16*	*1845*	*-0.13*	-8
Mueser, Troske, Gorislavsky (2007)	0.62	*1.6*	*45*	*1.3*	*38*	*1.07*	*1504*	*-0.28*	*-172*
LaLonde (1986), Dehejia, Wahba (1999)	0.44	*1.2*	*38*	*0.7*	*24*	0.03	*322*	*-0.90*	*-461*
Heckman, Smith (1999)	0.55	*1.3*	*44*	*0.9*	*31*	*0.60*	*952*	*-0.63*	*-404*
Dolton, Smith (2010)	0.38	*1.8*	*90*	*1.1*	*73*	*0.68*	*2751*	*-1.14*	*-375*
Baseline with very inflexible employment, unemployment & out-of-labour-force history	0.42	*1.2*	*53*	*0.6*	*39*	0.03	*1049*	*-0.97*	*-444*
Baseline with inflexible employment, unem-ployment & out-of-labour-force history	0.48	*1.3*	*57*	*0.9*	*49*	*0.69*	*2054*	*-0.48*	*-200*
		Training - women							
True model	1.00	-0.1	-2	-0.1	-2	-0.06	-58	0.00	3
Sianesi (2004)	0.83	*0.8*	*27*	*0.7*	*27*	*0.79*	*1289*	*-0.10*	*26*
Mueser, Troske, Gorislavsky (2007)	0.68	*1.7*	*30*	*1.7*	*30*	*1.36*	*1310*	*-0.26*	*-83*
LaLonde (1986), Dehejia, Wahba (1999)	0.50	*1.9*	*23*	*1.6*	*17*	*0.76*	*237*	*-0.27*	*-126*
Heckman, Smith (1999)	0.62	*1.6*	*24*	*1.6*	*24*	*1.41*	*1109*	*-0.21*	*-156*
Dolton, Smith (2010)	0.44	*1.7*	*68*	*1.6*	*70*	*1.52*	*3297*	*-0.54*	19
Baseline with very inflexible employment, unemployment & out-of-labour-force history	0.50	*1.0*	*35*	*0.9*	*35*	*0.47*	*1474*	*-0.51*	*-105*
Baseline with inflexible employment, unem-ployment & out-of-labour-force history	0.56	*0.9*	*35*	*0.9*	*38*	*0.82*	*1977*	*-0.23*	8
		Job search assistance - men							
True model	1.00	0.1	1	0.1	1	0.09	65	0.01	-1
Sianesi (2004)	0.91	*0.4*	*5*	*0.3*	4	*0.33*	*225*	-0.04	*-30*
Mueser, Troske, Gorislavsky (2007)	0.68	*1.0*	*18*	*0.6*	*14*	*0.28*	*414*	-0.02	3
LaLonde (1986), Dehejia, Wahba (1999)	0.42	*0.4*	*-8*	*-0.3*	*-21*	*-1.06*	*-1574*	*-0.45*	*-222*
Heckman, Smith (1999)	0.62	*1.8*	*11*	*0.7*	*-8*	-0.01	*-989*	*-0.27*	*-143*
Dolton, Smith (2010)	0.54	*0.7*	*-17*	*-0.4*	*-37*	*-1.24*	*-2443*	*-0.71*	*-319*
Baseline with very inflexible employment, unemployment & out-of-labour-force history	0.35	*1.4*	-2	-0.1	*-32*	*-1.38*	*-2682*	*-0.67*	*-342*
Baseline with inflexible employment, unem-ployment & out-of-labour-force history	0.42	*1.7*	*10*	*0.6*	*-12*	*-0.37*	*-1280*	*-0.13*	*-76*
		Job search assistance - women							
True model	1.00	-0.1	1	-0.2	0	-0.06	50	0.00	-5
Sianesi (2004)	0.89	0.1	*6*	0.0	*6*	-0.01	*264*	*0.17*	*109*
Mueser, Troske, Gorislavsky (2007)	0.64	*1.0*	*16*	*0.7*	*15*	*0.29*	*491*	-0.03	*67*
LaLonde (1986), Dehejia, Wahba (1999)	0.42	*0.3*	-4	0.0	*-9*	*-1.16*	*-1049*	*-0.13*	-9
Heckman, Smith (1999)	0.58	*1.2*	*3*	*0.7*	0	*0.20*	*-230*	*-0.10*	18
Dolton, Smith (2010)	0.52	*0.4*	*-5*	0.0	*-7*	*-0.43*	*-438*	*-0.70*	*-112*
Baseline with very inflexible employment, unemployment & out-of-labour-force history	0.40	*0.7*	*8*	*0.5*	4	*-0.72*	*-489*	*-0.34*	-18
Baseline with inflexible employment, unem-ployment & out-of-labour-force history	0.45	*1.1*	*15*	*0.8*	*12*	0.01	*268*	0.00	*102*

Note: *Italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level. Standard errors are obtained directly from the 500 simulation samples.

The results indicate that the specifications of all benchmark studies would lead to biased results. For training, the effects on employment and earnings would be overestimated and those on unemployment and UI benefit receipt would be underestimated. Overall, for training the specification by Dolton and Smith (2010) performs worst in most cases. The bias is relatively large and it is significant even in the actual data (see the Internet Appendix I.4, which contains all actual data results). For job search assistance, there is no worst specification, as the results very much vary with subsample and outcome variables. Interestingly, the LaLonde-type specifications perform surprisingly well for training of men, while the Sianesi-type specification, which has the propensity score with the highest correlation with the true propensity score, performs well for the training of women and job search assistance in general.

6.3 Does the specification really matter in applications?

The estimates of the bias and their lack of significance for the actual data (Table B.1 in Appendix B), as well as the corresponding Wald tests, which do not reject most specifications (see Internet Appendix I.4), may suggest that the bias from leaving out important variables is of no statistical relevance in applications that are of similar sample size as our study. We therefore assess whether the policy conclusions of the restricted models would be different from those based on the full model. The estimation results for all specifications with the actual data (see Internet Appendix I.4) show that the sign of the estimated effect differs very rarely. However, there are significance changes in a non-negligible number of cases, which would then lead to different policy conclusions. Moreover, the size of the effects differs considerably. In combination, this could have important implications for the results of cost-benefit analyses for the programs.

In Table 6.3, we perform a simple exercise to assess this problem. We count the number of specifications that differ from the full specification in terms of significance for each subsample and outcome. We use the 10% significance level as benchmark. For training, a large number of differences occur for cumulated earnings for men as well as for cumulated UI benefits for females. The problem is less severe for job search assistance of men, while for women the employment outcome and the cumulated UI benefits appear to be problematic.

Table 6.3: Estimated effects and differences in significance in actual data

		Training				Job search assistance			
		Men		Women		Men		Women	
		Effect	Fraction different in %	Effect	Fraction different in %	Effect	Fraction different in %	Effect	Fraction different in %
Employed 4 years after program start in %		*2.5*	27	*4.5*	7	-0.6	0	1.6	61
Half-monthly earnings	4 years after …	29	27	*84*	11	-66	23	13	9
Average employment	in year 4 after … in %	1.3	27	3.5	13	0.0	2	0.2	0
Average half-monthly earnings in year 4 after …		12	13	*85*	7	-56	13	-16	0
Cumulated employment	4 years after in half-months	*-4.5*	2	-0.8	5	*-4.8*	0	*-3.7*	5
Cumulated earnings	4 years after in half-months	-3181	46	1682	43	*-4145*	2	-2551	13
Cumulated UI receipt	4 years after in half-months	*-2.2*	0	*-1.0*	0	*-1.4*	0	*-1.0*	9
Cumulated UI benefits	4 years after …	*-670*	0	-102	41	*-594*	0	*-198*	73

Note: Effect refers to the effect estimated with the full model in the actual data. *Italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level. Standard errors are obtained from 499 bootstrap replications. Fraction different is the fraction of specifications in which the p-value is higher than 10% in case the p-value of the benchmark effect is at most 10%, and vice versa. The total number of specifications is 56.

Although more differences occur for the parsimonious specifications, several models that leave out only one bloc of variables are affected as well. Interestingly, the outcomes affected most are particularly demanding in terms of selection correction because several dimensions of labour market performance are affected: Firstly, the cumulated outcomes require balancing predictors for both short- and long-run performance. Secondly, given the fact that the conclusions for cumulated employment rarely change, balancing predictors of earnings seem to be particularly important. This is also true for cumulated UI benefits because they are a function of the previous earnings. These variables are also particularly important for cost-benefit analyses because they are the returns to program participation. The differences in significance in combination with large differences in the size of the effects lead to the conclusion that although estimated bias might not be statistically significant in applications, omitting important variables in the selection correction procedure may lead to wrong cost-benefit analyses and hence, wrong policy conclusions.

7. Conclusion

This paper investigates which groups of variables are required as control variables for classical evaluation studies of typical active labour market programs that rely on validity of the unconfoundedness, selection-on-observables or conditional independence assumption. We use a unique simulation design that ensures known true program effects, a realistic program assignment mechanism, and the validity of the unconfoundedness assumptions for the benchmark estimate in the data we use. Our results for typical European-style job search assistance and training programs indicate that very rich data is required to justify identification based on selection on observables.

We confirm the findings of the earlier literature in that controlling for caseworker assessments (Sianesi, 2004), pre-treatment outcomes (Mueser, Troske, and Gorislavsky, 2007), transitions between different labour market states and detailed regional information (Friedlander and Robins, 1995, Heckman, Ichimura, Smith, and Todd, 1998, Heckman and Smith, 1999) as well as for labour market histories in a flexible way (Dolton and Smith, 2010) is very important. However, we also find that information on the health of the unemployed worker and to some extent firm characteristics of the last employer, which have not been considered before, is important for selection correction. Regarding labour market histories, both short- and long-run histories play a role, as well as variables that cover multiple dimensions such as employment, unemployment, periods out of the labour force and earnings. Additionally, accounting for the timing of unemployment and program start as well as job search behaviour is relevant.

Complementing the simulation results with an analysis of actual data we find that leaving out one or more important blocs of variables has strong impacts on the inputs of cost-benefit analyses. Lack of important control variables may therefore lead to wrong policy conclusions regarding the cost-effectiveness of the programs.

Our results strongly suggest that in many countries further attempts to improve the information contained in the administrative data bases used to evaluate active labour market programs are required. However, this should go along with providing larger data bases as well, because, in a mean squared error sense, both sample size and informational content are equally important to obtain precise and reliable knowledge about the effects of these programs.

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, John, and Francis 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.

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

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

Dehejia, Rajeev H., and Sadek Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs", *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.

Dolton, Peter, and Jeffrey A. Smith (2010): "The Impact of the UK New Deal for Lone Parents on Benefit Receipt", mimeo.

Dorsett, Richard (2006): "The New Deal For Young People: Effect on the Labor Market Status of Young Men", *Labor Economics*, 13, 405–422.

Fitzenberger, Bernd, and R. Völter (2007): "Long-Run Effects of Training Programs for the Unemployed in East Germany", *Labour Economics*, 14, 730-755.

Fitzenberger, Bernd, and Stefan Speckesser (2007): "Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany", *Empirical Economics*, 32, 530-573.

Fraker, Thomas, and Rebecca Maynard (1987): "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources*, 22,195-226.

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.

Friedlander, D., and P.K. Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *The American Economic Review*, 85, 923-937.

Frölich, Markus, and Michael Lechner (2010): “Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies”, *Journal of the American Statistical Association*, 105, 1014-1029.

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

Heckman James J., and Jeffrey A. Smith (1999): "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program 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.

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

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

Heckman, James J., and V. Joseph 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, 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., Hidehiko Ichimura, and Petra E. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd (1998): "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-1098.

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 Labor Economics*, 3, 1865-2097, Amsterdam: North-Holland.

Heinrich, Carolyn J., Peter R. Mueser, Kenneth R. Troske, Kyung-Seong Jeon, and Daver C. Kahvecioglu (2009): “New Estimates of Public Employment and Training Program Net Impacts: A Nonexperimental Evaluation of the Workforce Investment Act Program”, IZA discussion paper 4569.

Huber, Martin, Michael Lechner, and Conny Wunsch (2010): "How to control for many covariates? Reliable estimators based on the propensity score", IZA discussion paper 5268.

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, 5-86.

Jacob, Brian A., Jens Ludwig, and Jeffrey Smith (2009): "Estimating Neighborhood Effects on Low-Income Youth", mimeo.

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

Joffe, Marshall M., Thomas R. Ten Have, Harold I. Feldman, and Stephen E. Kimmel (2004): "Model Selection, Confounder Control, and Marginal Structural Models", *The American Statistician*, November, 58, 272-279.

Khwaja, Ahmed, Gabriel Picone, Martin Salm, and Justin G. Trogdon (2010): "A Comparison of Treatment Effects Estimators Using a Structural Model of AMI Treatment Choices and Severity of Illness Information From Hospital Charts", forthcoming in the *Journal of Applied Econometrics.*

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

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

Larsson, Laura (2003): "Evaluation of Swedish Youth Labor Market Programs", *Journal of Human Resources* 38, 891-927.

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 Conny Wunsch (2009): "Active Labour Market Policy In East Germany: Waiting For The Economy To Take Off", *The Economics of Transition*, 17, 661–702.

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

Lechner, Michael, and Stephan Wiehler (2011): "Does the order and timing of active labor market programs matter?", *Oxford Bulletin of Economics and Statistics*, forthcoming.

Lechner, Michael, Ruth Miquel, and Conny Wunsch (2007): "The Curse and Blessing of Training the Unemployed in a Changing Economy: The case of East Germany after Unification", *German Economic Review*, 8, 468–507.

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

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

OECD (2010): OECD Employment Outlook 2010, Paris.

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

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

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

Sianesi, Barbara (2004), "An Evaluation of the Swedish System of Active Labor Market Programs 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.

Wunsch, Conny, and Michael Lechner (2008): "What did all the money do? On the general ineffectiveness of recent West German labour market programmes", *Kyklos: International Review for Social Sciences*, 134-174.

Appendix A: Further descriptive statistics and probit estimates

Table A.1 Further descriptive statistics and probit estimates for the actual data

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Constant				-0.182				0.025				*-0.920*				-0.247
Baseline characteristics																
Age in years	33.4	36.9	23.7	-0.006	34.3	37.8	23.5	*0.010*	35.2	36.7	10.4	-0.002	36.9	38.0	7.5	-0.005
Age 20-24 years	0.24	0.18	11.5	-0.107	0.20	0.15	8.5	-0.077	0.17	0.19	2.4	0.066	0.11	0.15	8.3	-0.050
Age >= 50 years	0.07	0.18	23.8	*-0.244*					0.10	0.18	16.2	*-0.258*				
Age 50-54 years					0.07	0.10	9.3	*-0.239*					0.08	0.10	6.4	-0.141
Age >= 55 years					0.01	0.10	29.1	*-0.726*					0.03	0.11	22.7	*-0.922*
No school degree	0.12	0.12	0.3	-0.023	0.07	0.08	3.5	-0.179	0.09	0.11	5.4	-0.050	0.03	0.08	14.4	-0.078
Upper secondary school degree	0.20	0.15	9.2	0.016	0.29	0.28	1.3	0.022	0.20	0.16	8.1	0.049	0.34	0.28	7.9	-0.058
University entry school degree	0.12	0.11	1.5	-0.012	0.17	0.16	1.8	0.022	0.18	0.11	15.0	0.029	0.21	0.15	10.9	-0.030
No vocational degree	0.37	0.34	3.8	-0.057	0.32	0.32	0.2	-0.065	0.27	0.33	10.1	0.039	0.22	0.33	17.2	0.022
University or college degree	0.03	0.04	2.2	*0.149*	0.05	0.05	1.3	-0.025	0.08	0.03	13.8	-0.016	0.06	0.05	5.3	0.018
No German citizen	0.15	0.17	3.2	-0.062	0.09	0.11	5.6	-0.159	0.13	0.15	5.3	-0.099	0.07	0.12	12.1	*-0.214*
At least one child	0.24	0.23	1.4	-0.013	0.38	0.33	8.6	*0.075*	0.23	0.23	0.1	*0.097*	0.41	0.32	13.0	0.056
At least one child < 3 years					0.04	0.03	1.6	*-0.174*					0.03	0.03	1.4	-0.112
At least one child 3-5 years					0.12	0.10	3.5	-0.017					0.15	0.11	9.3	0.013
Single	0.55	0.48	9.4	0.126	0.37	0.35	2.7	0.017	0.51	0.48	4.5	0.010	0.33	0.35	2.8	-0.087
Married	0.34	0.43	11.9	*0.159*	0.40	0.49	12.8	-0.074	0.40	0.43	3.9	-0.051	0.46	0.48	3.0	*-0.150*
Lone parent					0.11	0.07	9.1	0.036					0.11	0.08	7.0	-0.031

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Timing of entry into unemployment and program																
Beginning of unemployment	36.5	32.2	17.7	*-0.011*	34.6	31.7	11.7	*-0.024*	30.1	31.7	7.0	0.000	28.3	31.5	12.9	-0.011
Beginning of unemployment Dec-Feb	0.22	0.31	14.3	*-0.093*	0.20	0.21	2.5	-0.117	0.24	0.32	11.5	-0.062	0.20	0.22	3.1	-0.110
Beginning of unemployment Jun-Aug	0.28	0.24	6.3	-0.020	0.31	0.29	2.1	-0.047	0.26	0.24	2.2	0.019	0.29	0.29	0.3	-0.058
Beginning of unemployment Sep-Nov	0.27	0.25	4.3	-0.011	0.23	0.26	5.5	-0.053	0.26	0.24	3.5	0.043	0.25	0.26	1.7	-0.051
Time to treatment in half-months	6.77	5.25	19.3	*0.016*	6.9	5.5	18.1	0.012	8.0	6.5	18.5	*0.020*	7.8	6.3	20.6	*0.027*
Time to treatment 1	0.22	0.28	10.0	0.069	0.19	0.22	4.2	*-0.158*	0.09	0.11	5.9	*-0.062*	0.08	0.14	12.3	0.041
Time to treatment 2	0.12	0.14	5.2	*0.122*	0.14	0.16	5.1	-0.049	0.12	0.12	0.9	-0.036	0.11	0.14	6.0	0.052
Time to treatment 7-12	0.23	0.16	12.0	*0.128*	0.22	0.19	3.9	-0.012	0.29	0.25	6.1	*0.193*	0.26	0.23	5.8	0.052
Time to treatment > 12	0.18	0.12	13.1	0.165	0.19	0.11	16.7	-0.044	0.21	0.13	15.2	0.041	0.21	0.13	14.9	0.117
Last employment: non-firm characteristics																
Unskilled worker	0.41	0.37	5.2	0.029	0.23	0.21	2.9	-0.076	0.33	0.37	6.3	0.069	0.13	0.22	16.2	0.070
Clerk	0.18	0.16	3.6	*0.306*	0.35	0.35	0.3	0.039	0.31	0.16	26.5	0.099	0.50	0.35	21.3	-0.002
Part-time job					0.28	0.31	4.4	-0.038					0.28	0.31	4.7	-0.011
Occupation: Technical	0.18	0.16	2.5	0.088					0.21	0.16	9.0	-0.050				-0.148
Occupation: Construction	0.20	0.25	9.5	0.047					0.14	0.26	21.9	-0.001				-0.053
Occupation: Technical or construction					0.04	0.03	2.3	0.176					0.05	0.03	6.9	
Occupation: Service higher skilled	0.23	0.23	0.9	0.029	0.52	0.52	0.1	*0.242*	0.31	0.22	13.3	-0.042	0.67	0.51	22.8	-0.120
Occupation: Other	0.23	0.22	1.6	0.077	0.19	0.19	0.1	*0.000*	0.21	0.22	1.0	-0.109	0.13	0.19	10.9	0.000
Last employment: firm characteristics																
Age of firm	316	338	6.1	0.000	345	356	3.2	0.000	321	338	5.0	0.000	332	356	6.7	0.000
Firm size	269	321	1.9	*0.000*	233	270	2.6	-0.292	232	320	3.8	0.000	271	269	0.1	*-0.444*
Closed firm	0.10	0.09	2.4	*-0.369*	0.08	0.08	1.4	0.044	0.10	0.09	1.8	-0.248	0.10	0.08	3.4	-0.093
Fraction minor employees	0.09	0.09	1.0	0.033	0.13	0.13	2.2	0.008	0.08	0.09	0.6	-0.013	0.13	0.13	3.2	0.048
Fraction part-time employees	0.15	0.15	1.7	-0.043	0.28	0.30	3.3	0.001	0.15	0.15	0.1	0.000	0.27	0.30	7.1	*-0.008*
Mean age of employees	33.8	34.9	5.9	-0.004	34.7	35.5	4.4	-0.012	34.4	34.9	3.2	-0.006	34.8	35.5	4.0	-0.007

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Fraction of temporary workers	0.35	0.33	0.6	-0.008	0.27	0.32	2.0	0.000	0.30	0.34	1.8	-0.001	0.24	0.32	3.5	0.000
Mean tenure of employees	1234	1220	0.9	*0.000*	1289	1337	3.0	*0.218*	1317	1219	5.8	0.000	1345	1345	0.0	*0.138*
Fraction of non-German employees	0.27	0.25	5.1	*0.195*	0.24	0.23	2.2	-0.112	0.26	0.24	4.3	0.097	0.26	0.23	6.1	*-0.122*
Fraction of female employees	0.26	0.24	4.7	-0.025	0.57	0.59	3.8	0.040	0.27	0.24	7.5	0.075	0.54	0.59	11.9	0.031
Most frequent occupation: Technical	0.14	0.13	2.7	*-0.159*	0.05	0.05	1.6	0.078	0.16	0.13	5.6	-0.033	0.08	0.05	11.0	-0.024
Most frequent occupation: Construction	0.16	0.22	11.3	*-0.133*	0.02	0.02	0.7	-0.108	0.12	0.23	19.9	-0.048	0.04	0.02	5.0	0.045
Most frequent occupation: Service higher skilled	0.25	0.24	0.3	*-0.124*	0.45	0.44	1.4	*-0.102*	0.29	0.24	8.0	-0.055	0.48	0.44	5.9	0.042
Most frequent occupation: Other	0.23	0.21	3.4	-0.111	0.20	0.20	0.7	0.007	0.22	0.20	2.0	0.019	0.15	0.20	8.2	0.109
Industry: Retail	0.16	0.12	8.6	-0.069	0.21	0.19	4.2	0.058	0.16	0.12	8.5	-0.007	0.23	0.19	7.0	0.051
Industry: Financial services	0.16	0.13	5.6	-0.034	0.17	0.16	3.0	*-0.142*	0.19	0.13	11.8	-0.101	0.20	0.16	8.9	0.022
Industry: Education and health					0.16	0.17	3.2	0.002					0.13	0.17	8.9	0.175
Industry: Missing	0.05	0.07	3.4	-0.078	0.06	0.06	0.1	-0.041	0.06	0.07	2.1	*-0.373*	0.05	0.06	2.1	0.075
Industry: Construction	0.18	0.24	11.4	*-0.188*					0.14	0.25	19.8	*-0.148*				
Industry: Other services	0.11	0.12	2.0	*-0.222*					0.10	0.12	3.8	*-0.162*				
Industry: Other (men)	0.10	0.12	4.6	*-0.161*					0.10	0.12	5.4	*-0.165*				
Industry: Manufacturing					0.17	0.16	1.7	0.014					0.17	0.17	0.0	0.011
Industry: Other (women)					0.09	0.10	2.3	-0.013					0.11	0.10	3.0	-0.005
Short-term (2 years) labour market history																
Half-months employed in last 6 months	7.5	7.7	3.3	-0.030	8.2	8.0	2.2	-0.014	7.9	7.7	3.6	*-0.029*	8.5	8.0	7.3	-0.010
Half-months employed in last 24 months	30.4	30.2	1.1	-0.002	32.1	31.4	3.3	0.001	31.9	29.9	8.8	0.014	33.2	31.3	8.7	-0.002
Time since last employment if in last 24 months	4.9	4.6	1.8	-0.001	4.0	4.2	1.9	0.051	4.7	4.7	0.2	0.000	3.8	4.3	4.1	0.157
No employment in last 24 months	0.93	0.92	1.6	0.195	0.93	0.93	0.6	-0.002	0.94	0.92	5.2	-0.004	0.94	0.93	3.0	-0.051
Number of employers in last 24 months	1.78	1.69	4.9	-0.022	1.6	1.7	1.4	0.011	1.8	1.7	5.2	-0.039	1.6	1.7	3.3	0.031
Half-months unemployed in last 6 months	1.82	1.69	2.8	-0.014	1.3	1.4	3.2	-0.018	1.5	1.7	5.9	-0.009	1.1	1.5	9.6	-0.020
Half-months unemployed in last 24 months	10.1	10.3	1.0	0.001	6.8	7.9	6.9	0.004	8.7	10.5	10.6	0.005	5.8	8.1	14.4	-0.024
Unemployed 6 months before	0.26	0.25	1.5	-0.041	0.19	0.22	4.6	-0.033	0.22	0.26	5.5	-0.025	0.18	0.22	7.6	-0.054

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Time since last unemployment if in last 24 months	8.90	9.92	6.0	0.005	7.3	7.3	0.0	*0.371*	9.0	10.1	6.7	0.000	6.2	7.4	7.8	*0.228*
No unemployment in last 24 months	0.41	0.36	7.2	0.217	0.56	0.50	7.6	-0.077	0.45	0.34	15.5	0.125	0.63	0.50	18.9	-0.053
Unemployed 24 months before	0.24	0.28	7.1	-0.107	0.17	0.20	5.5	*0.008*	0.21	0.29	12.7	-0.034	0.14	0.20	11.7	0.006
Number of unemployment spells in last 24 months	0.65	0.78	9.9	-0.026	0.43	0.58	13.5	0.055	0.61	0.80	14.8	-0.007	0.39	0.59	17.8	-0.095
Any program in last 24 months	0.19	0.13	11.9	-0.116	0.14	0.11	6.3	0.033	0.16	0.13	5.3	0.041	0.13	0.11	4.7	0.071
Half-months out of labour force in last 6 months	2.2	2.1	1.6	-0.024	1.9	1.9	0.4	0.005	2.0	2.1	1.7	-0.005	1.8	1.9	2.7	0.000
Half-months out of labour force in last 24 months	6.7	6.8	0.5	-0.003	8.3	7.9	2.3	-0.020	6.7	6.8	1.3	0.003	8.1	7.8	1.9	-0.014
Out of labour force 6 months before	0.15	0.16	0.2	0.008	0.16	0.15	0.7	-0.081	0.15	0.16	0.7	-0.059	0.14	0.15	2.0	0.049
Out of labour force 24 months before	0.15	0.15	1.0	0.001	0.23	0.21	3.6	0.023	0.15	0.15	0.9	-0.022	0.24	0.21	4.6	-0.035
Time since last out of labour force if in last 24 months	6.0	6.0	0.1	-0.002	7.0	6.8	0.8	-0.002	5.9	6.0	0.4	0.001	6.5	6.8	2.1	0.000
No out of labour force in last 24 months	0.48	0.49	1.8	-0.049	0.48	0.49	0.4	-0.084	0.51	0.49	4.1	0.021	0.54	0.48	7.5	-0.047
Number of out of labour force spells in last 24 months	0.80	0.78	1.4	*-0.060*	0.72	0.75	2.3	-0.054	0.68	0.79	8.4	0.023	0.63	0.76	10.9	-0.022
	Long-term (10 years) labour market history															
Half-months employed in last 10 years	125	137	12.7	-0.001	120	133	14.1	0.000	135	137	1.7	-0.001	137	133	3.3	0.000
Tenure with last employer	22.6	22.8	0.4	0.000	26.2	22.6	6.9	-0.001	22.8	22.4	1.0	-0.001	21.9	22.5	1.2	0.001
Average employment duration	49.6	51.4	2.4	*0.001*	52.9	56.8	5.2	0.000	57.6	50.5	9.5	0.000	61.6	56.8	5.9	0.000
Number of employers in last 10 years	4.9	4.5	7.7	*0.014*	4.0	4.0	2.2	-0.004	4.8	4.5	4.9	0.010	3.8	4.0	2.9	0.021
Total time with last employer in last 10 years	47.6	62.5	16.6	0.000	55.6	64.9	10.1	0.000	52.3	61.9	10.3	-0.001	63.4	65.7	2.4	*0.001*
Half-months unemployed in last 10 years	34.9	35.3	0.8	-0.001	24.6	26.1	2.7	-0.001	31.3	35.6	7.5	-0.001	21.2	26.4	10.3	0.002
Time since last unemployment if in last 10 years	28.9	24.7	7.1	*0.001*	33.4	27.4	8.4	0.000	30.4	24.3	10.0	*0.001*	32.6	27.1	7.6	0.000
No unemployment in last 10 years	0.22	0.22	0.6	0.079	0.33	0.33	0.1	-0.062	0.26	0.21	8.3	0.034	0.40	0.33	10.3	-0.095
Number of unemployment spells in last 10 years	1.67	2.13	14.5	*-0.026*	1.0	1.4	13.6	-0.022	1.5	2.2	19.3	*-0.030*	0.95	1.36	15.4	*-0.056*
Duration of last unemployment spell	18.3	17.2	2.3	0.000	16.3	14.9	3.2	0.000	16.9	17.5	1.4	0.000	13.3	15.0	4.3	0.001
Average unemployment duration	15.3	14.0	3.7	0.000	12.4	12.1	0.8	0.000	14.0	14.1	0.3	0.000	11.1	12.1	3.4	0.000
Any program in last 4 years	0.23	0.16	12.3	0.110	0.17	0.13	6.7	0.094	0.21	0.16	8.3	-0.083	0.17	0.14	6.6	0.003
Any program in last 10 years	0.29	0.21	13.0	0.081	0.23	0.19	6.4	0.045	0.28	0.22	9.9	-0.002	0.23	0.19	7.5	-0.104

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Number of programs in last 10 years	0.44	0.29	13.9	0.046	0.32	0.25	7.6	0.061	0.38	0.30	8.8	*0.150*	0.32	0.25	6.7	*0.142*
Half-months out of labour force in last 4 years	16.6	15.5	4.1	-0.002	23.1	20.8	6.8	0.003	16.3	15.5	2.9	0.000	22.7	20.7	5.5	-0.001
Half-months out of labour force in last 10 years	77.4	65.3	13.0	0.001	94.0	78.9	15.8	0.001	71.0	65.5	5.9	0.000	80.6	78.3	2.4	0.002
Out of labour force 4 years before	0.28	0.24	6.2	0.055	0.39	0.33	9.2	0.001	0.27	0.24	4.5	0.024	0.34	0.33	1.5	0.107
Time since last out of labour force if in last 10 years	0.10	0.15	8.7	-0.062	0.09	0.12	7.4	0.110	0.14	0.14	0.7	-0.009	0.14	0.12	4.4	-0.012
No out of labour force in last 10 years	46.8	44.8	2.3	0.000	48.0	46.5	1.9	0.001	47.1	44.4	3.3	0.000	50.9	46.1	5.7	0.000
Number of out of labour force spells in last 10 years	2.77	2.58	5.9	*-0.035*	2.4	2.3	1.1	*-0.027*	2.4	2.6	6.5	-0.015	2.0	2.3	11.7	-0.014
Distance to hypothetical labour market entry	50.3	40.7	10.8	*-0.001*	58.4	46.7	12.1	0.000	43.5	41.0	2.9	0.000	44.6	46.2	1.8	0.000
Distance to hypothetical labour market entry non-Germans	11.4	10.4	1.8	0.000	8.2	7.6	1.3	0.001	9.2	9.5	0.6	0.000	5.8	7.8	4.5	0.001
Average out of labour force duration	36.5	31.8	7.3	-0.001	51.9	43.0	11.8	-0.001	35.9	31.8	6.3	-0.001	45.7	42.5	4.5	0.000
Benefits and UI claim																
Amount of unemployment benefit	311	311	0.0	-0.010	223	214	5.6	0.000	3.2	3.1	2.0	0.000	2.2	2.1	3.9	0.000
Remaining UI claim	276	315	13.8	*0.000*	302	332	11.2	0.000	308	315	2.5	0.000	335	333	0.5	0.000
No UI claim	0.14	0.12	4.5	*-0.257*	0.08	0.08	0.8	-0.113	0.09	0.12	6.4	-0.091	0.06	0.08	5.5	0.034
UI claim 1-5 months	0.11	0.11	0.5	*-0.180*	0.09	0.12	6.2	*-0.244*	0.09	0.11	4.6	*-0.170*	0.08	0.12	10.5	-0.080
UI claim 6-8 months	0.15	0.16	1.0	-0.011	0.19	0.17	4.9	*-0.198*	0.18	0.16	4.6	*-0.159*	0.14	0.17	4.4	0.017
UI claim 9-11	0.13	0.13	1.4	*-0.080*	0.10	0.12	3.5	*-0.175*	0.12	0.14	3.1	*-0.110*	0.09	0.12	6.4	-0.084
UI claim > 12 months	0.13	0.20	15.1	0.041	0.15	0.22	12.8	-0.052	0.17	0.20	6.5	-0.036	0.19	0.22	5.6	0.002
Compliance with benefit conditions, employability and mobility																
Fully mobile within Germany	0.35	0.34	1.8	0.017	0.39	0.37	3.3	-0.028	0.38	0.34	6.1	-0.005	0.38	0.37	1.7	0.026
Average number of vacancy referrals	0.09	0.11	1.8	-0.050	0.07	0.08	0.9	-0.025	0.08	0.11	2.9	0.039	0.06	0.08	2.1	0.090
No vacancy referral	0.16	0.34	29.8	*-0.368*	0.17	0.36	32.1	*-0.251*	0.18	0.33	24.1	*-0.526*	0.22	0.36	21.8	*-0.578*
Any form of non-compliance with benefit conditions	0.24	0.19	9.2	0.052	0.11	0.10	2.9	-0.038	0.19	0.19	0.0	*0.084*	0.07	0.10	7.2	0.001

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Health																
Health impairment or disability	0.17	0.22	9.0	-0.069					0.16	0.22	11.3	0.005				
Health impairment or disability affects employability	0.10	0.14	10.0	-0.086					0.09	0.15	11.9	-0.049				
Health impairment					0.15	0.21	11.3	-0.031					0.14	0.22	15.3	-0.043
Health impairment affects employability					0.08	0.12	9.4	*-0.118*					0.06	0.13	15.1	0.040
Disability					0.03	0.05	6.7	-0.040					0.03	0.05	8.3	-0.071
Total duration reported in sick during receipt of benefits	1.4	1.6	2.9	-0.002	1.2	1.4	3.6	0.001	1.3	1.6	6.7	*-0.012*	1.1	1.4	6.6	-0.009
Did not report in sick during receipt of benefits	0.69	0.69	0.4	-0.003	0.74	0.73	1.5	-0.036	0.72	0.68	5.7	-0.072	0.75	0.72	4.4	-0.040
Characteristics of job looked for																
Looking for high-skill job	0.04	0.05	2.8	0.019	0.05	0.06	2.5	-0.029	0.09	0.04	13.6	-0.072	0.07	0.05	6.0	-0.059
Looking for unskilled or skilled job	0.45	0.43	3.1	-0.005	0.41	0.40	1.9	-0.015	0.35	0.42	10.4	*0.061*	0.29	0.41	17.9	0.053
Occupation looked for: Technical	0.21	0.19	4.0	*0.082*					0.26	0.19	12.7	*0.079*				
Occupation looked for: Construction	0.22	0.26	7.2	**-0.115**					0.14	0.26	21.5	0.047				
Occupation looked for: Technical or construction					0.04	0.03	6.0	*0.277*					0.06	0.03	9.4	*0.259*
Occupation looked for: Service higher skilled	0.22	0.21	1.4	0.014	0.53	0.53	0.3	*0.201*	0.29	0.21	13.7	0.014	0.70	0.52	25.2	0.012
Occupation looked for: Other	0.20	0.19	2.5	-0.058	0.20	0.19	0.8	-0.029	0.16	0.18	4.7	0.063	0.11	0.19	15.8	*0.088*
Looking for part-time job					0.27	0.28	1.6	-0.016					0.34	0.28	8.5	-0.054
Detailed Regional information																
Regional GDP growth	19.8	20.9	6.7	-0.001	19.8	20.7	6.1	*0.003*	20.2	21.0	4.8	0.002	21.2	20.7	2.5	0.001
Travel time to next big city on public transport	61.3	66.9	8.5	0.000	60.3	64.4	6.4	0.000	64.5	67.4	4.5	0.000	61.4	64.5	4.8	0.000
Share of non-Germans in region	9.1	9.2	1.3	0.002	9.0	9.5	7.5	0.009	9.3	9.2	1.2	*-0.008*	9.7	9.5	2.7	*-0.010*
Local unemployment rate in %	8.8	8.3	11.5	-0.007	8.5	8.2	6.8	-0.005	8.5	8.3	4.5	*0.014*	8.2	8.2	0.4	0.016
Big city	0.48	0.47	2.7	*-0.126*	0.49	0.48	0.8	0.049	0.46	0.47	0.5	-0.074	0.52	0.48	5.9	0.019
Rural area	0.10	0.15	11.9	*0.073*	0.11	0.13	5.3	0.030	0.13	0.16	5.0	*-0.076*	0.11	0.13	4.3	-0.068
Net migration	3.4	3.8	9.6	-0.005	3.5	3.9	8.2	*-0.013*	3.8	3.8	1.5	*-0.024*	3.8	3.9	1.3	*-0.023*

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
	Pre-treatment outcomes															
Employed 4 years before	0.56	0.56	0.3	0.054	0.51	0.54	4.9	0.104	0.58	0.56	3.6	*0.158*	0.57	0.54	3.2	*0.209*
Earnings 4 years before	786	910	9.2	0.000	564	627	6.0	0.000	920	900	1.5	0.000	669	625	4.2	0.000
Cumulated duration employed 4 years before	59.4	60.2	2.1	-0.001	59.1	59.9	2.2	-0.001	61.8	60.0	4.7	*-0.004*	61.3	59.9	3.7	-0.002
Cumulated earnings 4 years before	52.1	57.3	9.3	*-0.004*	38.2	39.9	3.8	0.010	60.7	56.8	6.5	0.002	44.0	39.7	9.3	0.004
Cumulated duration of UI 4 years before	7.8	9.9	13.5	0.000	5.9	7.5	11.5	0.003	7.4	10.0	17.1	0.001	5.6	7.7	14.6	0.006
Cumulated UI benefits 4 years before	1.5	2.0	16.3	-0.011	0.81	1.10	12.4	-0.070	1.4	2.1	17.7	-0.015	0.82	1.12	12.8	*-0.039*
	Region dummies															
Baden-Wurttemberg	0.12	0.12	0.3	-0.045	0.13	0.14	2.6	0.116	0.12	0.11	0.9	0.058	0.15	0.14	1.7	0.116
Bavaria	0.09	0.23	28.0	*-0.160*	0.12	0.21	17.4	0.017	0.15	0.23	14.5	*-0.320*	0.17	0.21	7.0	-0.107
Lower Saxony, Bremen	0.17	0.16	1.5	*0.072*	0.15	0.15	0.4	0.120	0.19	0.16	4.7	0.019	0.15	0.15	1.0	*0.087*
Schleswig-Holstein, Hamburg	0.19	0.07	25.2	*0.205*	0.20	0.08	25.4	*0.327*	0.11	0.07	9.6	*0.550*	0.11	0.07	8.9	*0.625*
Hessen	0.07	0.08	2.3	-0.052	0.07	0.08	3.4	-0.018	0.08	0.08	0.0	-0.067	0.07	0.08	2.1	-0.017
Rhineland- Palatinate, Saarland	0.08	0.08	0.8	0.002	0.07	0.07	1.1	*0.328*	0.08	0.08	0.2	0.018	0.11	0.07	9.2	0.093
	Industry- and occupation-specific experience															
Average duration in last occupation	25.1	24.7	0.8	-0.001	28.2	24.6	6.5	0.000	24.5	24.1	0.8	-0.001	23.9	24.3	0.9	0.000
Average duration in last industry	20.5	19.6	2.9	*-0.003*	22.0	19.4	7.8	-0.002	18.6	19.2	2.0	0.000	18.3	19.4	3.4	0.000
Total duration in last occupation	75.4	93.0	16.7	-0.001	82.8	97.1	13.9	0.000	80.1	92.7	11.7	0.000	97.5	97.7	0.2	0.000
Total duration in last industry	40.6	42.0	2.8	0.001	41.8	41.5	0.5	0.000	39.6	41.7	4.1	0.002	39.6	41.4	3.8	0.000
Number of occupations in last 10 years	3.7	3.3	10.4	0.011	2.9	2.7	5.8	0.016	3.6	3.3	8.3	0.011	2.7	2.7	2.5	0.007
Number of industries in last 10 years	2.2	1.9	12.9	0.030	2.0	1.9	7.6	-0.005	2.1	1.9	10.2	*0.022*	1.8	1.8	0.9	0.024
	Earnings history															
Earnings in last job	833	867	5.3	0.001	599	603	0.5	0.000	9.4	8.6	10.9	0.000	6.7	6.0	10.7	0.000
Average earnings in last 10 years	661	723	14.4	0.011	517	537	5.7	0.000	7.5	7.2	6.0	0.000	5.8	5.3	10.9	0.000
Cumulated earnings in last year	14592	15042	2.8	0.005	11293	10887	3.1	0.000	17.1	14.9	13.0	0.000	13.0	10.8	15.8	0.000
Cumulated earnings in last 2 years	14170	15528	8.2	0.004	10429	10882	3.3	0.000	16.8	15.4	8.2	*0.000*	12.1	10.8	9.5	*0.000*

Table A.1 to be continued

Table A.1 Further descriptive statistics and probit estimates for the actual data (continued)

	Job search assistance								Training							
	Men				Women				Men				Women			
Variable	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff	P	NP	SD	Coeff
Outcomes																
Employed 4 years after	0.45	0.41	5.8		0.51	0.43	10.2		0.49	0.41	10.4		0.52	0.43	13.5	
Earnings 4 years after	786	799	0.8		598	531	5.9		961	785	11.1		708	510	16.2	
Cumulated duration employed 4 years after	34.2	39.0	10.3		36.7	38.8	4.5		36.7	38.4	3.5		40.3	38.0	4.7	
Cumulated earnings 4 years after	34653	41034	11.3		25038	27091	4.6		41397	40104	2.1		31641	25961	12.2	
Cumulated duration of UI 4 years after	11.3	15.9	26.0		11.2	15.2	22.5		10.9	15.5	25.9		11.5	15.1	19.7	
Cumulated UI benefits 4 years after	4316	6534	25.6		3040	4003	16.1		4588	6436	20.3		3347	4001	10.6	

Note: P: Mean among participants (fractions if not stated otherwise), NP: Mean among nonparticipants (fractions if not stated otherwise), SD: Absolute standardized difference in percent (difference in sample means of respective participants and corresponding nonparticipants divided by the square root of the sum of the empirical variances in the two subsamples). Coeff: Estimated coefficient of a probit model for selection into the respective program. The probit models also include several interaction terms between the beginning of unemployment and time to treatment, as well as time to treatment and vacancy referrals.

Reference groups for dummies are omitted. 'before' and 'after' means before and after the beginning of the unemployment spell that determines membership in our population of interest. If not mentioned otherwise, all variables are measured at the beginning of this unemployment spell. Variables related to information in this spell are measured at the (simulated for controls) start of the program. Earnings are measured as earnings per half-month. 'Cumulated' measures sum up the half-monthly measures. Beginning of unemployment spell is measured in half-months where the first half of January 2000 equals '1'. All monetary measures are in EUR of the year 2000.

Appendix B: Further estimation results

Table B.1: Regression results for the estimations based on actual data

	4 years after program start		Average in year 4 after program start		Cumulated effects over the first 48 months after program start			
	employ-ment rate in %	half-monthly earnings in EUR	months employed in%	half-monthly earnings in EUR	half-months employed	earnings in EUR	half-months on UI	benefit receipt from UI in EUR
Timing of entry into unemployment & program	*0.71*	7.4	0.20	3.5	0.04	36	-0.17	-68
Last job: Non-firm characterist.	-0.07	1.4	-0.06	4.6	-0.08	178	-0.02	-13
Firm characteristics	-0.02	-4.4	-0.36	**-10.4**	*-0.36*	*-489*	*-0.14*	*-69*
Labour market history: 2 years	0.01	-5.1	-0.48	-9.1	-0.43	-551	-0.21	*-95*
10 years	0.42	4.1	0.02	-1.2	-0.03	-319	-0.06	*-69*
Earnings history	0.18	*10.3*	0.30	*9.5*	0.27	311	0.08	16
Industry- & occupation-specific experience	0.16	-6.1	-0.04	-7.4	0.00	-192	-0.08	-24
Pre-treatment outcomes	-0.10	-4.2	0.02	-2.4	-0.17	-157	0.01	29
Benefits & UI claim	-0.22	1.9	-0.10	2.6	-0.12	-63	0.09	62
Compliance with benefit condit., employability & mobility	-0.09	-4.8	-0.22	-6.5	-0.03	-150	0.04	-7
Health	*0.47*	**9.9**	**0.45**	***12.5***	**0.50**	**587**	*0.20*	82
Characteristics of job looked for	0.02	1.6	0.17	2.8	-0.05	-149	-0.12	-66
Region dummies	-0.09	-2.4	-0.28	-7.3	-0.27	-469	-0.04	-46
Detailed regional information	0.07	-1.8	-0.18	-4.5	-0.22	-219	-0.07	-23
History: Employment	0.18	8.8	0.35	9.8	-0.03	282	-0.05	20
Unemployment	-0.48	-6.0	-0.19	-8.3	-0.35	*-545*	-0.02	-38
Out-of-labour-force	0.12	5.9	0.20	11.1	0.36	763	0.17	99

Note: The entries refer to the coefficients of a regression of the bias (estimated program effect minus effect estimated using the full model) on dummies that equal one if the respective bloc of variables is left out in the estimation of the propensity score. *Italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level. Standard errors are obtained from 499 bootstrap replications. Sample size: 200 observations.

Appendix C: Technical details of the matching estimator used

Table C.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: $\sum_{i=1}^{N} \frac{\underline{1}(D=1)\hat{y}^0(x_1)}{N^1} - \frac{\underline{1}(D=0)w_i\hat{y}^0(x_0)}{N^0}$. 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. It 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'.

For the estimations based on the actual data, there is an issue on how to draw inference. Abadie and Imbens (2008) show that for matching estimators with a fixed number of comparison observations bootstrap-based inferences are 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, we use the bootstrap. It is implemented following MacKinnon (2006) by bootstrapping the p-values of the t-statistic directly based on symmetric rejection regions. 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.

An additional (Internet) Appendix can be found in the discussion paper versions published in the series of the University of St. Gallen, the CEPR, and the IZA.