Sensitivity of matching-based program evaluations to the availability of control variables

Michael Lechner[+] and Conny Wunsch[++] [*]

[+] Swiss Institute for Empirical Economic Research (SEW) of the University of St. Gallen

[++]Department of Economics of the VU University Amsterdam

Revised: January 2013

Date this version has been printed: ***08 June 2016***

Highlights:

- **Selection of covariates affects results in evaluation studies of active labour market programs**
- **Paper shows the extent of this sensitivity and identifies a key set of variables**
- **Paper also shows how previous studies may be affected by a lack of particular covariates**
- **Method used is Empirical Monte Carlo Study based on very informative German administrative data**

[*] Michael Lechner is also affiliated with CEPR and PSI, London, CES-Ifo, Munich, IAB, Nuremberg, and IZA, Bonn. Conny Wunsch is also affiliated with CES-Ifo, Munich, IZA, Bonn, and Tinbergen Institute, Amsterdam. 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 was 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, at a workshop on education at the University of Barcelona and at a workshop on labour market evaluation at Harvard University. We are grateful for the comments we received in these seminars and further discussions. We particularly thank Alberto Abadie, Sergio Firpo, Steve Lehrer, Edwin Leuven, Jeff Smith, and Chris Taber. The usual disclaimer applies.

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 about the current unemployment spell, regional information, pre-treatment outcomes, and detailed short-term labour market histories remove most of the selection bias.

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

JEL classification: J68

Addresses for correspondence: Michael Lechner, Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, Switzerland, Michael.Lechner@unisg.ch, www.michael-lechner.eu. Conny Wunsch, VU University Amsterdam, Department of Economics, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, c.wunsch@vu.nl.

1. Introduction

This paper contributes to the literature that investigates the question which variables are needed to correct for selection bias when evaluating active labour market programs for unemployed workers. Most of the existing studies use experimental data from the U.S. and compare the results obtained from the experimental control group with the results from a non-experimental control group.[1] Besides the usual concerns about experiments summarized in Heckman and Smith (1995), an important drawback of the data used in these studies is that they do not contain the wealth of information argued to be required for credible selection correction in non-experimental data.[2] Hence, it is unclear whether the findings from this literature can be generalized to existing active labour market programs with selective entry of participants.

We address this issue by using German administrative linked employer-employee data that contains information on all major factors claimed to be important for selection correction and were used in applications that rely on the unconfoundedness assumption. We analyze job search assistance and training programs in West Germany that are the most widely used programs in OECD countries and typical in terms of their contents, implementation, and selection of participants. We employ a simulation design that is based on actual data proposed by Huber, Lechner, and Wunsch (2013). By simulating placebo participants from the actual non-participants, this design ensures that the true effect is known to be zero, that the selection model is known and that the unconfoundedness assumption holds. Moreover, in contrast to common Monte Carlo studies, it imposes no assumptions about the relation between covariates and outcomes but exploits their dependencies as they are in the data which is crucial for assessing the value of available covariate information.

Our analysis confirms the findings from the earlier literature that basic socio-demographic information together with pre-treatment outcomes (Mueser, Troske, and Gorislavsky, 2007), re-

[1] 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), Smith and Todd (2005): see also Heckman and Hotz (1989), Peikes, Moreno, and Orzol (2008), and Shadish, Clark, and Steiner (2008).

[2] 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 (2011) for Germany, Jespersen, Munch, and Skipper (2008) for Denmark and Heinrich, Mueser, Troske, Jeon, and Kahvecioglu (2009) for the USA.

gional information and labour market histories (Friedlander and Robins, 1995, Heckman, Ichimura, Smith, and Todd, 1998, Heckman and Smith, 1999, Dolton and Smith, 2010) are important for removing biases. More information than suggested in these studies is required, though, when evaluating typical European-style programs. In particular, certain information on the unemployment spell as well as more detailed short-run labour market histories are needed in addition to remove most of the bias. We also provide an easy-to-use tool to predict potential biases that might arise in similar applications based on different data. Applying this tool to specifications used by past evaluation studies of European programs, we find no quantitatively important biases. In particular, the most recent studies provide useful benchmarks for future applications of this type.

The paper proceeds as follows: The next section outlines the research design. In Section 3, we provide details on the data, the programs analyzed, the estimation sample and the matching estimator used. Section 4 analyses selection into the programs. Section 5 presents the results. The last section concludes. An Appendix as well as an additional Internet Appendix provides further details on the data, the estimator, the results and the robustness checks.

2. Empirical design

We employ the simulation design proposed by Huber, Lechner, and Wunsch (2013) which simulates a placebo treatment the true effect of which is known to be zero. We estimate a selection model using actual participants and nonparticipants in the program (as would be done in an application). Then, based on the predicted participation probability, a fraction of actual nonparticipants that is equal to the observed fraction of participants is assigned randomly to the placebo treatment. Hence, the coefficients of the model estimated from actual participants and nonparticipants become the 'true' selection parameters in the placebo data. This ensures, firstly, that the true selection model is known and as close as possible to real selection decisions. Secondly, the unconfoundedness assumption holds by construction in the simulated data. External validity requires that the estimated selection model on which the simulations are based captures all major confounders and, because we are interested in the average effect on the treated (ATET), that actual and placebo participants are similar. These issues are addressed in Section 4. The advantage with respect to classical Monte Carlo studies is that, because we use the actual data of nonparticipants, we do not need to specify

data generating processes for the confounders, the outcomes, or their interaction. The latter is crucial for our research question.[3]

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

3. Data and implementation details

3.1 Data

We use a unique linked employer-employee administrative database. It is a 2% random sample drawn from the population of all German employees subject to social insurance since 1990, which covers 85% of the German workforce. It combines information from different administrative sources: (1) the records provided to social insurance by employers for each of their employees (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 jobseeker 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 also reliable, because it is used for counselling, job referral, monitoring, and assessing jobseeker's compliance with search requirements.

[3] Using a similar approach, Jacob, Ludwig, and Smith (2009) study applications for randomly assigned housing vouchers using the fact that for randomized out applicants the effect of the voucher is zero. 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 (2011) 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.

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) based on the population of all employees of the firm as of June 30 of each year in which the firm existed for 1990 to 2006 (so-called *Betriebshistorikpanel*). 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 particular spell, type of employment (full/part-time, high/low-skilled), occupation, earnings, type and amount of UI benefit, and the remaining potential UI benefit duration. It also includes information on compliance with the benefit conditions (e.g. failure to show up at interviews, refusal to participate in an assigned labour market program, imposition of sanction), and periods when a UI recipient reported sickness to the UI. Further, 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 training programs. The jobseeker register contains a wealth of individual characteristics, including date of birth, gender, educational attainment, marital status, number of children, age of youngest child, nationality, profession, the presence of health impairments, and disability status. With respect to job search the data contain the type of job looked for (full/part-time, high/low-skilled, occupation), whether the jobseeker is fully mobile within Germany and whether she has health impairments that affect her 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.

The German administrative linked employer-employee data used here is probably the most comprehensive dataset currently available for the evaluation of typical job search assistance and

training programs for the unemployed. Clearly, administrative data outperform usual survey data 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) for the U.S., 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.[4]

Table 1: Relation of our dataset to other administrative datasets

Study	Country	Variables missing compared to our data	Variables missing in our data
Gerfin and Lechner, 2002)	Switzerland	firm characteristics, less informative in terms of health and job search	maternity leave
Lechner and Wiehler, (2012)	Austria	firm variables, less informative in terms of health and job search	subjective caseworker assessment of the employability of each jobseeker
Sianesi (2004, 2008)	Sweden	health, marital status, number and age of children, occupation and skill profile of the last job, firm characteristics of the last job other than industry, occupation looked for, employment histories	caseworker's assessment of the client's job readiness, need for guidance and difficulty to be placed
Jespersen, Munch, and Skipper (2008)	Denmark	occupation and skill profile of the last job, firm characteristics, anything related to job search, several dimensions of labour market histories, in particular long-run histories	
Lechner, Miquel, and Wunsch (2007, 2011); Fitzenberger and Speckesser (2007); Fitzenberger and Völter (2007); Fitzenberger, Osikominu, and Völter (2008)	Germany	health, anything related to job search, firm characteristics other than industry and firm size*	
Lechner and Wunsch (2009); Wunsch and Lechner (2008)	Germany	firm characteristics other than industry**	

Note: * Information on remaining UI claims is also not available but can be calculated from the data. This has been done by Lechner, Miquel, and Wunsch (2007, 2011) but not in the other papers mentioned. ** Although this information is available in previous versions of the German administrative data, occupation and industry-specific experience beyond the last job have also not been considered in any of the mentioned studies.

In Table 1, we compare a number of studies that use selection-on-observable strategies to identify program effects based on quite informative administrative data to our data. We report both

[4] So far, linked employer-employee data have been used mainly for other labour market analysis than the evaluation of labour market programs (see Abowd and Kramarz, 1999).

the variables that are missing in these studies compared to our data, and variables that are available in the respective study but not ours. 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 will be argued below in Section 4.1, 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 Internet Appendix I.2, we construct a large variety of different measures from the data to capture all important aspects of individual labour market histories.[5]

3.2 Programs considered

We focus on the two types of programs that are most widely used in developed economies: job search assistance and training. The former comprise the typical combination of counselling, referral to vacancies, monitoring, and job search training in Germany (e.g. Thomsen, 2009). With respect to training, we focus on the internationally most typical programs that provide vocational training and have planned durations of at most six months.[6] The implementation of the programs we look at is also largely representative with respect to eligibility and selection into the programs as we describe in detail in Section 4.

3.3 Sample selection and definition of participation status

Since we are interested in typical programs in economies beyond the transition stage, we exclude East Germany (and Berlin). We start with an inflow sample into unemployment in the period April 2000 to December 2002.[7] We restrict the analysis to the population aged 20-59 in order to avoid having to model educational choices or (early) retirement decisions. We ensure eligibility for program participation by requiring all individuals to qualify for unemployment insurance benefits or

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

[6] 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).

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

unemployment assistance (UI), which is paid (subject to a means test) after exhaustion of UI. We also exclude (a few) individuals who start their unemployment spell with a program or for whom the information from the jobseeker register is missing.

As in Lechner, Miquel, and Wunsch (2011), we define as (non-) participants all those individuals who (do not) start a program within the first 12 months of their unemployment spell.[8] 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 distributions of start dates (job search assistance is used very early in the spell while training starts later). 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 the simulated program start.[9] 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.[10]

3.4 Outcome variables

We consider eight outcome variables that measure different dimensions. All outcome variables are measured relative to the start of the program.[11] The majority of studies report employment

[8] Nonparticipation means not starting any program in the 12-month window, not just the program used for the particular comparison. We subjected the 12-month window to a sensitivity check by also considering 6- and 18-month windows. The results remain qualitatively the same. The details of this sensitivity check are contained in Internet Appendix I.10 in Table I.21.

[9] Related to the arguments of Fredriksson and Johansson (2003, 2008), Sianesi (2004), and Lechner and Wiehler (2012) 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.

[10] This implies that individuals who become unemployed in late 2002 will be under-sampled as they are likely to have their real or simulated program starts after December 2002.

[11] Some studies measure outcome variables relative to the end of the program. However, as program duration is usually endogenous, using the beginning of the program has become the norm in the more recent studies.

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.

3.5 Estimation

We focus on the average treatment effect on the treated (ATET). We use a matching estimator because 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). As is common in the literature, we tackle the dimensionality problem by conditioning on a parametric estimate of the conditional participation probability (so-called propensity score, see Rosenbaum and Rubin, 1983) rather than on the selection variables directly. We use binary probit models for the propensity score. The full specification in the actual data, that includes all blocks listed in Table 2 and is used to simulate placebo participants, has been tested extensively against misspecification (non-normality, heteroscedasticity, omitted variables).[12] To ensure common support, we discard participants who have a higher or lower propensity score estimate than, respectively, the maximum or minimum in the comparison group. Only one participant in job search assistance and two training participants were removed. There was also sufficient overlap in the other parts of the distributions of the propensity score.[13] After this step,

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

[13] 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 to the training program, because those observations would never appear in any match.

the propensity score was re-estimated on the common support.[14] There are no common support issues in the simulated data either.

We use the matching estimator suggested by Lechner, Miquel, and Wunsch (2011) because it is one of the best estimators investigated in a simulation study by Huber, Lechner, and Wunsch (2013). It incorporates the idea of caliper or radius matching (e.g. Dehejia and Wahba, 2002) combined with weighted regressions that use the weights obtained from matching. This procedure reduces small sample bias as well as asymptotic bias (see Abadie and Imbens, 2006) and increases robustness. Internet Appendix I.1 describes the details of this estimator. To assess match quality we checked the means and medians of a large number of potential confounders in the treated and matched control samples. The after-match balance is high for all comparisons.

4. Selection into the programs

4.1 Do we observe all relevant factors in this study?

Our research design guarantees validity of the selection-on-observables assumption in the placebo data (internal validity). However, external validity, i.e. the question whether what we do can be generalized to other studies that use actual participants, is only achieved if the selection model used for the simulation is plausible, i.e. if it includes all relevant confounders. We exploit that due to budget and capacity constraints not everyone who is eligible, satisfies the selection criteria and is willing to participate can get a program slot. Next, we discuss the factors that determine program participation.

Eligibility requires receipt of UI benefits or unemployment assistance and is ensured by the way the sample is constructed (see Section 3.3). According to German legislation, caseworkers have to select job seekers based on local labour market conditions, employment prospects and qualification needs. To measure local labour market conditions we observe the rich set of regional indicators listed in Section 3.1 that allow controlling for the relevant regional differences in a detailed way. The determinants of employment prospects are captured by personal characteristics like age, gen-

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

der, marital status, nationality, number of children, and age of youngest child. Skills are measured in terms of schooling and vocational training as well as by 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 are calculated from 10 years of pre-unemployment employment histories.

Unobserved heterogeneity in motivation, productivity, and employability is captured indirectly in several ways: Firstly, we use 10 years of detailed labour market histories to control for the quality and stability of employment, the frequency and duration of previous unemployment experience, and other periods of non-employment. Secondly, we condition on the characteristics of the last employer that may reveal specific types of workers. Thirdly, we control for incidence of non-compliance with benefit conditions during past unemployment spells. Fourthly, 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 worker's employability. Finally, we know whether the jobseeker is fully mobile within Germany.

In addition to the measures of skills, productivity, experience, and motivation that were already mentioned, we account for the type of job previously held as compared to the one looked for in terms of full/part-time, high/low-skilled and occupation to determine potential qualification needs. For job search assistance, we capture potential job search experience and job search skills by past unemployment experience and their average duration.

As program assignment is the outcome of decisions made by both caseworker and jobseeker there is also self-selection of unemployed workers into the programs. Firstly, similar to many other countries, there are institutional incentives to participate. Jobseekers refusing to participate in a program they were assigned to risk a benefit sanction. Moreover, for our period of investigation, and this is a feature mainly of some European countries, participation in training (but not in job search assistance) stops the clock for exhausting UI benefits. Since there are also benefit payments during the program, jobseekers effectively extend their potential UI benefit duration by participating in

training. We directly capture these incentives by controlling for the amount of benefits, UI eligibility and remaining potential UI benefit duration.

Table 2: Blocks of control variables

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

Note: The details of these variables are contained in Internet Appendix I.2. All variables are measured prior to the unemployment spell that is used to define participation and non-participation.

The second set of factors that drive self-selection are 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

(2011) show that program participation and becoming pregnant during unemployment may be competing 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. 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 10-year labour market histories. 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 children and the age of the youngest child, is likely to capture aspects of fertility decisions and child raising preferences.

In summary, our data enable us to capture the main confounding factors that affect both program participation and labour market outcomes. Thus, the selection-on-observables assumption appears to be credible. Table 2 summarizes the blocks 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 3.5). 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 men. Therefore, all estimations are conducted separately for men and women (as well as for training and job search assistance). To improve the external credibility of our results we also conducted some robustness checks that are described in detail in Section 5.5.

4.2 Selectivity of program participation in the data

Internet Appendix I.2 presents the sample means of all variables that are used to estimate the 'true' propensity scores (full model) for participants and non-participant in each program in the actual data. We also report their absolute standardized difference in percent in order to assess the magnitude of potential selection bias as proposed by Imbens and Wooldridge (2009), as well as the estimated coefficients from the probit models. The main insights 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 selection is stronger for training than for job search assistance: For the latter 6-8% of all variables that are included in the true selection model show a standardized

difference above 15%, while for training the respective fraction is 10-11%. 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.

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. Therefore, based on the actual data we conduct Wald tests of the joint significance of the elements of the 15 blocks of variables (defined in Table 2) in the propensity-score and the outcome equations for both programs considered. For the outcome equations, we estimate probit models for binary outcome variables and linear models for all other outcome variables (see Section 3.4) in the population of nonparticipants (since we are interested in the ATET). It is important to note that the purpose of the outcome regressions is purely descriptive to assess broadly the relevance of the blocks of variables. They are not an attempt to estimate the correct model and to derive causal conclusions.

The results, which are reported in Internet Appendix I.4, indicate that all blocks of variables we consider are strongly related 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 block of variables conditional on all other blocks being included in the model. Thus, if we leave out one of the other blocks, these blocks may become important nevertheless. Therefore, we keep them. Overall, the low p-values indicate strong statistical relevance for each individual block even given all the other blocks, implying that leaving them out is likely to bias results.

Given that we are interested in the ATET, external validity, i.e. the question whether the results we obtain for the placebo participants generalize to actual participants, not only requires observing all relevant confounders, but also that actual and placebo participants are comparable. Internet Appendix I.3 reports the descriptive statistics for all variables in the simulated data. The

comparison with Internet Appendix I.2 shows that actual and placebo participants are very similar in all characteristics, implying that the simulation results are informative for the actual participants.

5. Results

The following subsections summarize the results from 73 different specifications of the propensity score model. 69 of them differ in which of the 15 blocks of variables in Table 2 are included. The remaining four specifications mimic the specifications proposed in other studies. The tables provided in Internet Appendix I.5 show the full list of specifications.[15] To simplify the exposition we only discuss the results for men and women in training. All conclusions also hold for job search assistance and the results are available in the Internet Appendix. For the same reason, we also focus the discussion on two outcome variables: the average employment rate in the fourth year after program start as the most important longer-run outcome, and cumulated earnings over the four years after program start because it also captures potential lock-in effects and dynamics over time. We only report and discuss the biases of the ATET that occur when omitting blocks of variables from the true model. Although looking at the variance is interesting as well, we consider the bias to be the quantity most relevant for researchers, also because the variance can be estimated in an application while the bias cannot.

5.1 Hierarchical analysis of the relevance of particular blocks of variables

Table 3 shows the biases that occur for training when the selection model is extended sequentially, starting with the baseline, by adding blocks of variables. The order of adding variables should resemble the ease, likelihood or cost of obtaining the respective information for researchers. It is however not fully sequential, since, based on a set of variables included, we also investigate the effects of adding the components of the next block of covariates to understand which component might be the most important one (e.g. earnings versus employment history). To get some idea about the relation of the variables included to the variables still missing, in the second and third column of

[15] We do not vary the variables within blocks because of computation time. In total, we estimated 73 specifications, 500 times, in four subsamples, on both the simulated and the actual data, which adds up to 292,000 runs of the matching procedure.

these tables we present the correlations of the propensity score of the full model with the propensity score of the model used for that particular specification.[16]

Table 3: Bias of effects of training programs when sequentially adding covariates

				Outcome variables			
Specification of propensity score		Correlation of *p-scores* in %		Average employment rate in year 4 in %		Cumulated earnings after 4 years in EUR	
		Men	Women	Men	Women	Men	Women
Baseline		36	45	*0.6*	*1.6*	213	*1601*
Baseline characteristics and ...							
Timing of program start		54	61	*1.0*	*1.9*	*1215*	*1632*
Region dummies		42	47	*1.1*	*1.5*	*1581*	*1818*
Benefits and UI claim		39	51	*0.6*	*1.1*	*164*	*1774*
All of them		61	70	*0.8*	*1.2*	*896*	*1178*
Baseline characteristics, timing, region, UI benefit claim and ...							
Pre-program outcomes		64	71	*0.8*	*1.4*	*716*	*1045*
Baseline characteristics, timing, region, UI benefit claim, pre-program outcomes and ...							
Employment history	(last 2 years)	66	73	*0.9*	*1.1*	*965*	*834*
Unemployment history	(last 2 years)	67	76	*1.0*	*1.4*	*1414*	*1444*
Out-of labour-force history	(last 2 years)	65	74	*0.5*	*0.7*	*551*	*648*
Earnings history	(last 2 years)	66	74	*0.8*	*0.5*	*384*	-8
Non-firm characteristics	(last job)	73	74	*0.8*	*1.4*	*709*	*657*
All of them		78	82	*0.5*	*0.6*	*475*	*442*
Baseline characteristics, timing, region, UI benefit claim, pre-program outcomes, short-term labour market history and ...							
Employment history	(last 10 years)	79	83	*0.5*	*0.6*	*612*	*247*
Unemployment history	(last 10 years)	81	84	*0.4*	*0.5*	*688*	*434*
Out-of labour-force history	(last 10 years)	79	84	*0.5*	*0.5*	*621*	*229*
Earnings history	(last 10 years)	78	82	*0.4*	*0.6*	*562*	*439*
All of them		84	87	*0.4*	*0.5*	*516*	*208*
Baseline characteristics, timing, region, UI benefit claim, pre-program outcomes, short and long-term labour market history and ...							
Firm characteristics (last job)		89	92	*0.4*	*0.4*	*402*	185
Industry- and occupation-specific experience		86	88	*0.4*	*0.5*	*515*	*287*
Health		85	87	0.2	0.1	201	22
Compliance with benefit conditions, employability, mobility		92	91	*0.4*	0.0	*714*	27
Characteristics of job looked for		85	89	*0.4*	*0.4*	*417*	*134*
Detailed regional information		86	88	*0.4*	*0.6*	*305*	30
All of them = true model		100	100	-0.1	-0.1	-30	-58
Program effects and averages of outcome variables for the actual participants							
Effect of program participation		-	-	1.3	3.5	-3181	1682
Average of outcome variable (treated)		-	-	43	47	41397	31641

Note: *Correlation of p-scores* measures the correlation of the propensity scores of the full model and the reduced model indicated in the particular line for men and for women, respectively. *Italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level. The numbers appearing in the upper part of this table show the estimated biases that are obtained from the simulations (standard errors are obtained directly from the 500 simulation samples), while the estimated effect (standard errors are obtained directly by 499 bootstrap simulations) and the average outcome levels for the participants are based on the full sample and the actual data.

[16] The true model is also included as a consistency check. 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 (2013).

To obtain a better understanding of the economic relevance of the biases discussed below, it is important to get an idea about the magnitudes involved. Therefore, the last two rows in Table 3 report the effect of the programs as well as the means of the outcome variables for the actual program participants. From the latter we see that a bit less than half of them worked in year 4. In the four years after program start they earned between 31'000 and 41'000 EUR in total. Compared to these levels the program effects are small if existent at all. For men the estimated effect on employment in year four is tiny and insignificant, while the effect on cumulated earnings over four years is negative and significant due to the well-known lock-in effect, and it amounts to about 7.5% of the outcome level. For women, the cumulated earnings effect of training is small and statistically insignificant, but the employment effect in year 4 is positive and significant, and again about 7.5% of the respective outcome level.[17]

In the following, we quantify the biases that result from omitting covariates, relative to these magnitudes. A natural starting point is the baseline scenario, in which we only control for basic socio-demographic characteristics (see Table 2 for details). The first observation is that there is considerable gender heterogeneity. For men, the bias for all outcome variables is small and not much affected by the inclusion of further information. Bias even increase in some cases suggesting that adding more information is not necessarily better when there is still some information missing.[18] For women the bias is much larger though and ignoring it would lead to a substantial overestimation of the positive effects of training, in particular for the cumulated outcomes that also cover the lock-in effect. However, after additionally controlling for the variables related to the timing of the program, region, information on the benefits, pre-treatment outcomes and short-term labour market history, the remaining biases are small and similar for men and women. Although they could be reduced further, they have reached a level of 1-2% of the average level of the outcome variables so that it may not be worth investing additional resources to reduce them further. Thus, the conclusion from Table 3 is that controlling for this information seems to be enough to remove all

[17] Overall, these results are consistent with what has been found by Wunsch and Lechner (2008) in a similar setting.

[18] Heckman and Navarro (2004) also provide examples for how adding more information may lead to higher bias.

economically relevant biases. This is not surprising given the correlations of the respective propensity score with the propensity score of the full model. While for the baseline specification this correlation is only around 40%, after controlling for the variables mentioned it attains a level of about 80% leaving not much room for further improvements.

Our results are consistent with the previous literature: Heckman and Smith (1999) show the importance of regional information, while Mueser, Troske, and Gorislavsky (2007) point to the necessity of including pre-treatment outcomes. With respect to short-term labour market histories Table 3 shows that it is not sufficient to control only for certain aspects of it. Biases are reduced to the small levels reported above only if information on all relevant dimensions (employment, unemployment, out-of-labour-force status, earnings, non-firm characteristics of the last job) is included in a flexible way. This supports the conclusions from Heckman and Smith (1999) and Dolton and Smith (2010). We discuss the relation of our results to these studies in Section 5.3.

5.2 Regression analysis of the relevance of particular blocks of variables

The hierarchical approach employed in the previous section suggests a readily understandable stopping rule for the gathering of information but depends on the ordering of the blocks of variables considered. In specific applications, researchers may have access to different information sets and may want to understand what the likely bias might be if they use them, or how this bias could be reduced if some other specific information is added. One way to provide such information is to perform the above simulations for all possible combinations of blocks of variables. However, it is impossible to report the results from such an exercise in any accessible way, and computation time would be prohibitive. As an alternative, we conduct a regression analysis based on a subset of possible combinations of blocks of variables and assess its usefulness in predicting bias. The regressions are based on 69 specifications of the propensity model where we add to the baseline specification or leave out from the full model the blocks of variables described in Table 2, either individually or group wise. The data on which the regressions are based (i.e. the biases of all specifications considered) are presented in Internet Appendix I.5.

The regression results are presented in Table A.1 in the Appendix. The regression models are specified such that there is a constant term as well as a dummy variable indicating whether a specific block of variables is included in the estimation of the propensity score. Consequently, the constant term measures the scenario where only the baseline characteristics are included. The coefficients on a given dummy variables indicate by how much the bias changes relative to the baseline scenario by adding the respective block of variables. For example, the coefficient of the constant term in the regression for women in training means that the baseline specification exhibits a bias in the estimated employment effect of 1.5 percentage points. The coefficient on health means that including health information is predicted to reduce this bias by 0.6 percentage points. The predicted bias for any combination of the blocks of variables can be obtained by adding the corresponding coefficients to the coefficient of the constant term. For example, the bias in the employment effect when including benefit information and the two-year employment history on top of the baseline characteristics is predicted to be 0.6 percentage points for women in training.

The usefulness of Table A.1 for predicting bias depends on the accuracy of the linear approximation. The R^2's given in the last row of Table A.1 suggest that the fit is not perfect. Also, they are, with one exception, much higher for women than for men. To check the quality of the approximation in more detail we reproduce the predictions of Table 2 in Internet Appendix I.7. We compare estimated and predicted bias for the baseline model, all major groups of variables (rows 'All of them' in Table 2), as well as the full model. We find that there may be some larger deviations when considering the baseline scenario only, in particular for men. However, as soon as the variables related to the timing of the program, the region dummies, and the information on the benefits are included, the predictions from the regressions and the hierarchical analysis are close to each other and support the stopping rule that emerged from the hierarchical analysis. Thus, conditional on this relative small set of variables which should be available in almost all applications, Table A.1 provides a useful tool to assess potential biases in similar applications with different data. Note, however, that this only tests for the in-sample validity of the regression models since the results in Table 3 are part of the data used. In the next section, we provide some supporting evidence for out-of-sample validity.

5.3 Comparison with specifications proposed in other studies

Continuing this line of thought, we also assess potential biases resulting from specifications proposed in the literature, and how these potential biases may affect policy conclusions. We consider four benchmark studies. All of them have considerably gaps in information in several dimensions (see Section 3.1), but they emphasize specific types of control variables: 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 (see also the criticism by Smith and Todd, 2005). Heckman and Smith (1999) emphasize the importance of accounting for transitions between employment, unemployment, and out-of-labour-force status as well as regional differences. Mueser, Troske, and Gorislavsky (2007) point to the importance of including pre-treatment outcomes. Dolton and Smith (2010) advocate the necessity to control for pre-treatment outcomes in a sufficiently flexible way. [19]

To assess the proposed specifications, we proceed in two steps. Firstly, we use Table A.1 to predict biases. We sum-up the coefficients of the blocks of variables that broadly cover the types of variables used in the studies. However, the exact variables used there can deviate quite substantially from those included in the blocks of variables we consider. Therefore, in a second step, we run an additional set of simulations where we specify the propensity score models such that they mimic the information sets used in those studies as closely as possible (see Internet Appendix I.8 for the exact specifications used). Table 4 presents the results in a similar way as Table 3.

The first important finding from Table 4 is that despite the differences in the specifications, the predictions from Table A.1 and the direct simulations lead to similar results. For training and job search assistance, the correlation of the obtained biases is, respectively, 90% and 94% (all outcomes as well as men and women pooled). The very high correlation of the predicted and estimated

[19] We include this study although it looks at a different type of program and population. They use administrative data to evaluate Britain's New Deal for Lone Parents which is a program targeted at lone parents that provides information, referrals and limited financial support. The data cover age, region, age of youngest child, number of children, disability status, duration of disability, and benefit history. The latter correspond to the pre-treatment outcomes in their studies. The equivalent in our application is the unemployment history.

biases strongly supports the usefulness of Table A.1 for predicting biases in other applications and datasets, even when exact mimicking of the variables included in each block is impossible.

Table 4: Bias of effects for selected specifications obtained from simulations

			Outcome variables			
Specification of propensity score	Correlation of *p-scores* in %		Average employment rate in year 4 in %		Cumulated earnings after 4 years in EUR	
	Men	Women	Men	Women	Men	Women
LaLonde (1986), Dehejia, Wahba (1999)	44	50	*0.7*	*1.6*	*322*	*237*
prediction from Table A.1: blocks 0,1,6a,8a			0.3	1.0	154	696
Heckman, Smith (1999)	55	62	*0.9*	*1.6*	*952*	*1109*
prediction from Table A.1: blocks 0,2,6,8a,14			0.5	1.4	1036	1234
Mueser, Troske, Gorislavsky (2007)	62	68	*1.3*	*1.7*	*1504*	*1310*
prediction from Table A.1: blocks 0,1,2,4,6,8a			0.6	1.2	1105	1027
Dolton, Smith (2010)	38	44	*1.1*	*1.6*	*2751*	*3297*
prediction from Table A.1: blocks 0,2,6b,11			0.6	0.9	1076	1559
Program effects and averages of outcome variables for the actual participants						
Effect of program participation full model	-	-	1.3	3.5	-3181	1682
Average of outcome variable (treated)	-	-	43	47	41397	31641

Note: *Correlation of p-scores* measures the correlation of the propensity scores of the full model and the reduced model indicated in the particular line for men and for women, respectively. *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.

Table 4 consistently indicates that the effects on employment and earnings would be overestimated.[20] Thus, studies based on specifications like those may very well find positive effects of training programs while the true effect is zero or close to zero, leading to overly optimistic conclusions about the effectiveness of training. Consequently, claims that the specifications advocated in the main references used by applied researchers suffice to correct for selection bias (e.g. by Mueser, Troske, and Gorislavsky, 2007, for the JTPA program in the U.S. state Missouri) do not generalize to the evaluation of typical European labour market programs. It is important to note, though, that our results confirm the importance of the aspects of the specifications that have been emphasized in these papers (see the discussion in Section 5.1). We add the qualification that although they are important, they are not sufficient. In particular, none of the benchmark studies includes all of the blocks of variables that have been identified as crucial for reducing bias to a quantitatively negligible level (blocks 0-6 and 8a in Table 2) in Section 5.1.

[20] Analogously, the effects on unemployment and UI benefit receipt would be underestimated. See Internet Appendix I.9.

5.4 Implications for the conclusions from existing evaluation studies

In this section, we address the question whether our results have implications for what we know about the effectiveness of active labour market programs by assessing potential bias in existing evaluation studies. We focus on the main applications that evaluate European training or job search assistance programs for unemployed workers employing propensity-score matching. The reason is that for this type of study the external validity of our results is largest because they closely resemble our setup. In Sections 5.2 and 5.3, we have shown that the regression results presented in Table A.1 have sufficient validity to be a useful tool for out predictions. Therefore, we use this table to predict the bias existing evaluations might be subject to and discuss whether this may change overall conclusions.

The applications we consider are those listed in Table 1 where we show the relation of our data to the data used in other studies. They cover five different European countries and six different administrative datasets. For Germany (DE) we include Fitzenberger, Osikominu, and Völter (2008) and Lechner, Miquel, and Wunsch (2011) both using the first version of the German administrative data but with different specifications, as well as Wunsch and Lechner (2008) who use the predecessor of our data. Gerfin and Lechner (2002) and Lechner and Wiehler (2012) are included for Switzerland (CH) and Austria (AT), respectively. The Swedish study (SE) of Sianesi (2004, 2008)[21], as well as the Danish study (DK) by Jespersen, Munch, and Skipper (2008) are considered as well. For each of these applications we assess which blocks of variables in Table 2 are covered by the types of variables used in the studies. Then, we add the corresponding coefficients from Table A.1 to obtain a prediction of the potential biases. The results are displayed in Table 5.

Somewhat more sizeable biases only appear in two cases: The largest ones occur for Sianesi (2004, 2008) who also uses the smallest set of control variables. Our results imply that the effects of training reported by her are likely to be overestimated by 3-5% of the mean level of the outcome. They are larger by 1-2% of the mean outcome for men than for women. The biases stem from the

[21] Sianesi (2008) evaluates training and other Swedish programs separately while Sianesi (2004) considers all programs together. Both studies use the same specification of the propensity score, though.

lack of information on remaining UI claims, pre-treatment outcomes and several dimensions of (short-term) labour market histories which have been identified as most important in Section 5.1. However, they do not change the main conclusions of Sianesi (2004, 2008) because she finds negative effects. The results in Table 5 suggest that labour market training in Sweden may have been even more harmful than documented earlier.

Table 5: Predicted bias for selected applications

	Average employment rate in year 4 in %		Cumulated earnings after 4 years in EUR	
	Men	Women	Men	Women
DE: Fitzenberger, Osikominu, Völter (2008): 0,1,2,5,6a,7a,8,9,14	0.8	1.2	193	11
Lechner, Miquel, Wunsch (2011): 0,1,2,3,4,5,6,7,8,14	0.4	0.6	120	3
Wunsch, Lechner (2008): 0,1,2,3,4,5,6,7,8a,11,12,13,14	-0.3	-0.3	-211	-3
CH: Gerfin, Lechner (2002): 0,1,2,3,5,6a-b,7a-b,8a,12,13,14	0.8	0.4	350	9
AT: Lechner, Wiehler (2012): 0,1,2,3,5,6,7,8a,9,11,14	-0.1	0.3	463	11
SE: Sianesi (2004, 2008): 0,1,2,5,6b,7b,8a,12	1.3	1.0	951	25
DK: Jespersen, Munch, Skipper (2008): 0,1,2,3,5,6a,7a,11,14	0.1	0.3	128	10
Program effects and averages of outcome variables for the actual participants				
Effect of program participation (for actual treated)	1.3	3.5	-3181	1682
Average of outcome variable (for actual treated)	43	47	41397	31641

Note: The biases are predicted by adding the relevant coefficients reported in Table A.1. The blocks of variables, the coefficients of which are added, are indicated by the corresponding numbers after the reference. The estimated effects and the average outcome levels for the actual participants are the ones reported in Table 2. For the estimated effects for actual participants: *italics*: significant on the 10% level, **bold**: significant on the 5% level, ***bold italics***: significant on the 1% level.

For men (but not for women) biases of 2-3% of mean outcomes are predicted for Gerfin and Lechner (2002) who use Swiss data. As the study finds no effects of training, our results imply that in fact they might have been (slightly) negative. Yet this would not change policy conclusions because the programs are not cost-effective anyway. The reason for the biases is that labour market histories are captured only relatively crudely. It is interesting though, to compare their specification with Jespersen, Munch, and Skipper (2008), for whom no quantitatively important biases are predicted. In the Danish study, labour market histories are captured even more crudely but they observe some information on health, which is lacking in the Swiss data. As can be seen from Table A.1, health makes the largest individual contribution to the reduction of bias from the baseline specification. The comparison of the two studies suggests that information on health might help to compensate for lack of detailed information on short-term labour market histories.

The general message from Table 5 is that in the majority of cases predicted biases are substantively minor and that none of the conclusions from the main European evaluation studies changes. Interestingly, predicted biases are higher for the earlier studies reflecting improvements in the available data as well as some learning. The available information in the most recent studies seem to have reached a quality that can be regarded as a useful benchmark for future applications of this type that are based on comparably informative data.

5.5 Sensitivity analysis

In the following, we address potential concerns about the external validity of our results. The details are contained in Internet Appendix I.10 and I.11. Here we summarize the main points.

First, one may worry that the quality of the control variables is lower for foreigners than for Germans. This would raise concerns about the validity of the conditional independence assumption for this group of unemployed workers. One reason is that educational degrees may have been obtained in a foreign country and that translating this back into the categories of the German education system may be prone to considerable measurement error. Furthermore, language skills may be important for foreigners, but we do not have a good measure for this in our data. To address this issue we performed the key simulation for Germans only (baseline and full model). We obtain the same qualitative results as for the full sample.

The second issue relates to the definition of the treatment window of 12 months. This may seem somewhat arbitrary but should not affect the relative performance of different specifications because all are based on the same definition. We re-estimate the baseline and full model for treatment windows with a length of 6 and 18 months. The estimated biases of the 18-month window are somewhat more similar to those obtained for the 12-month window than those of the 6-month window. However, none of the differences is in an order of magnitude that changes conclusions.

Finally, to provide supporting evidence for the validity of the unconfoundedness assumption in the actual data and for the external validity of the selection model used to generate the simulated data, we performed a pre-program test in the spirit of Heckman and Hotz (1989). We artificially shifted the beginning of unemployment and hence (simulated) program start four years into the past for the actual data. We then use our identification and estimation strategy to estimate the program

effects for the four years up to the actual beginning of unemployment. As control variables, we only use history information preceding the four years before the actual beginning of unemployment. If our selection (correction) model is correct, the estimated program effects should be zero in this pre-treatment period because treated and controls should be comparable before program start. Internet Appendix I.10 shows the half-monthly estimates of the program effects in the four-year pre-treatment period for employment and earnings. There are no significant effects for the vast majority of half-months. Thus, this robustness check strongly supports our chosen approach.

6. 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 the validity of the selection-on-observables. We use a 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 rich data are required to justify identification based on selection on observables. From the magnitudes of the biases incurred, we conclude that basic socio-demographic information together with certain information on the unemployment spell, region and pre-treatment outcomes as well as detailed short-run labour market histories appear to be sufficient to remove most of the biases. Adding more information, like long-term labour market histories, health or job search information, may help to reduce the biases further, but only by a small amount. Hence, it may not be worth to invest extra resources in gathering such additional information. Note, however, that other variables, in particular health measures, may become important when some of the other information we identified as crucial is unavailable.

Our findings are consistent with the earlier literature which concludes that controlling for pre-treatment outcomes (Mueser, Troske, and Gorislavsky, 2007), transitions between different labour market states and regional information (Friedlander and Robins, 1995, Heckman, Ichimura, Smith, and Todd, 1998, Heckman and Smith, 1999), as well as labour market histories in a flexible way (Dolton and Smith, 2010) is very important. However, more information than suggested in these studies is needed for evaluating typical European-style programs. Their specifications would

lead to an overestimation of the effects of training and to an underestimation of the effects of job search assistance, at least in countries that use assignment mechanism which are similar to the ones used in Germany.

We also provide an easy-to-use tool to predict potential bias in applications that are based on different data. This can be useful for other researchers who would like to assess whether identification based on the unconfoundedness assumption is a viable option with the data they have in a similar type of application. We apply this tool to several existing evaluations of European job search assistance and training programs. The results indicate robustness of the conclusions drawn from this literature regarding the effectiveness of the programs. Furthermore, our findings suggest that the most recent studies provide useful benchmarks for future applications of this type.

References

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

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", *The Economic Journal*, 120, F452-F477.

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

Fitzenberger, Bernd , Aderonke Osikominu, and Robert Völter (2008): " Get Training or Wait? Long-Run Employment Effects of Training Programs for theUnemployed in West Germany", *Annals of Economics and Statistics*, 91/92, 321-355.

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

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.

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 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., and Salvador Navarro-Lozano (2004) "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models", *The Review of Economics and Statistics*, 86(1), 30-57.

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 Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies", *Economic Journal*, 109, 313-348.

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 (2013): "The Performance of Estimators Based on the Propensity Score", *Journal of Econometrics*, forthcoming.

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

Khwaja, Ahmed, Gabriel Picone, Martin Salm, and Justin G. Trogdon (2011): "A Comparison of Treatment Effects Estimators Using a Structural Model of AMI Treatment Choices and Severity of Illness Information From Hospital Charts", *Journal of Applied Econometrics*, 26(5), 825-853.

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, 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.

Lechner, Michael, Ruth Miquel, and Conny Wunsch (2011): "Long-Run Effects of Public Sector Sponsored Training in West Germany", *Journal of the European Economic Association*, 9(4), 742-784.

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 (2011): "Kids or Courses? Gender Differences in the Effects of Active Labor Market Policies", *Journal of Population Economics*, 24(3), 783-812.

Lechner, Michael, and Stephan Wiehler (2012): "Does the order and timing of active labor market programs matter?", *Oxford Bulletin of Economics and Statistics*, forthcoming.

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.

Peikes, Deborah N., Lorenzo Moreno, and Sean Michael Orzol (2008): "Propensity score matching", *American Statistician*, 62(3), 222-231.

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

Shadish, William R., M. H. Clark, Peter M. Steiner (2008): "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments", *Journal of the American Statistical Association*, 103(484), 1334-1344.

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.

Sianesi, Barbara (2008), " Differential effects of active labour market programs for the unemployed", *Labour Economics*, 15, 370-399.

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.

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

Appendix: Regression results for the bias

Table A.1: Regression results for the bias: training

	Average employment rate in year 4 in %		Cumulated earnings after 4 years in EUR	
Block of variables added to baseline specification	Men	Women	Men	Women
Baseline (constant in regression) (0)	*1.1*	*1.5*	*1094*	*1607*
Timing of entry into unemployment & program (1)	*-0.3*	*0.0*	*-290*	*-323*
Region dummies (2)	*0.1*	*0.0*	*518*	*57*
Benefits & UI claim (3)	*-0.2*	*-0.5*	*-468*	*-215*
Pre-treatment outcomes (4)	*0.1*	*0.0*	13	*-249*
Non-firm characteristics of last job (5)	*0.2*	*0.0*	-18	*-407*
Labour market history, 2 years, employment (6a)	*-0.2*	*-0.4*	*-161*	*-241*
unemployment (6b)	*0.2*	*0.0*	*453*	*258*
out-of-labour force (6c)	-0.1	0.2	-33	*265*
earnings (8a)	*-0.3*	*-0.1*	*-489*	*-347*
10 years, employment (7a)	0.1	0.0	*214*	55
unemployment (7b)	*0.2*	*0.0*	*435*	*200*
out-of-labour force (7c)	*-0.3*	*-0.3*	*-425*	*-215*
earnings (8b)	*-0.1*	-0.1	*-443*	*-284*
Firm characteristics (last job) (9)	*0.3*	0.1	*478*	*174*
Industry- & occupation-specific experience (10)	*0.1*	*0.2*	*219*	*207*
Health (11)	*-0.8*	*-0.6*	*-989*	*-363*
Compliance with benefit condit., employability, mobility (12)	*0.2*	*-0.5*	*286*	*-94*
Characteristics of job looked for (13)	*-0.2*	*0.0*	*-180*	*-157*
Detailed regional information (14)	0.0	*0.2*	-56	-42
R^2	33	69	41	74

Note: The entries refer to the mean - across simulations - of the coefficients of an OLS regression of the bias (equal to the estimated effect because the true effect is zero) on dummy variables that are equal to one if the respective block of variables is included 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: 69 observations. Standard errors are obtained directly from the 500 simulation samples. The numbers in parentheses in column one indicate to which block in Table 2 this corresponds.