Universität St.Gallen

Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach

Michael C. Knaus, Michael Lechner, Anthony Strittmatter

August 2017 Discussion Paper no. 2017-11

School of Economics and Political Science,
Department of Economics

University of St.Gallen

Editor: Martina Flockerzi
University of St.Gallen
School of Economics and Political Science
Department of Economics
Müller-Friedberg-Strasse 6/8
CH-9000 St. Gallen
Phone +41 71 224 23 25
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Müller-Friedberg-Strasse 6/8
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: http://www.seps.unisg.ch

Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach[1]

Michael C. Knaus[2], Michael Lechner[3], Anthony Strittmatter

Author's address:

Michael C. Knaus
Swiss Institute for Empirical Economic Research (SEW)
Varnbüelstr. 14
CH-9000 St. Gallen
Phone +41 71 224 2304
Email Michael.knaus@unisg.ch
Website www.sew.unisg.ch/

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW)
Varnbüelstr. 14
CH-9000 St. Gallen
Phone +41 71 224 2814
Email Michael.lechner@unisg.ch
Website www.michael-lechner.eu

Anthony Strittmatter
Swiss Institute for Empirical Economic Research (SEW)
Varnbüelstr. 14
CH-9000 St. Gallen
Phone +41 71 224 2305
Email Anthony.strittmatter@unisg.ch
Website www.anthonystrittmatter.com

[1] Financial support from the Swiss National Science Foundation (SNSF) is gratefully acknowledged. The study is part of the project "Causal Analysis with Big Data" which has grant number SNSF 407540_166999 and is included in the Swiss National Research Programme "Big Data" (NRP 75). A previous version of the paper was presented at the University of Maastricht, Department of Economics, at the workshop on unemployment and labour market policies, Barcelona, the IZA summer school in labor economics, Ammersee, the IAB, Nuremberg, and at the Computational Social Science Workshop, Konstanz. We thank participants, and in particular we thank Hugo Bodory, Bruno Crépon, Chris Hansen, Jeff Smith, and Martin Spindler, for helpful comments and suggestions. The usual disclaimer applies

[2] Michael C. Knaus is also affiliated with IZA, Bonn.

[3] Michael Lechner is also affiliated with CEPR and PSI, London, CESIfo, Munich, IAB, Nuremberg, and IZA, Bonn.

Abstract

We systematically investigate the effect heterogeneity of job search programmes for unemployed workers. To investigate possibly heterogeneous employment effects, we combine non-experimental causal empirical models with Lasso-type estimators. The empirical analyses are based on rich administrative data from Swiss social security records. We find considerable heterogeneities only during the first six months after the start of training. Consistent with previous results of the literature, unemployed persons with fewer employment opportunities profit more from participating in these programmes. Furthermore, we also document heterogeneous employment effects by residence status. Finally, we show the potential of easy-to-implement programme participation rules for improving average employment effects of these active labour market programmes.

Keywords

Machine learning, individualized treatment effects, conditional average treatment effects, active labour market policy.

JEL Classification

J68, H43, C21.

1 Introduction

In this study, we employ machine learning methods for a systematic investigation of effect heterogeneity of job search programmes ('JSPs' from now on) in Switzerland. Programme evaluation studies widely acknowledge the possibility of effect heterogeneity for different groups. Stratifying the data in mutually exclusive groups or including interactions in a regression framework are two baseline approaches to investigate effect heterogeneity (see, e.g., Athey and Imbens, 2017a, for a review). However, these approaches may overlook important heterogeneities because they usually do not include a *systematic* search based on clear, spelled-out statistical rules. Furthermore, for large-scale investigations of effect heterogeneity, standard p-values of classical (single) hypothesis tests are no longer valid because of the multiple-hypothesis testing problem (see, e.g., Lan et al., 2016, List, Shaikh, and Xu, 2016). For example, for fifty single hypotheses tests, the probability that at least one test falsely rejects the null hypotheses at the 5% significance level could be up to 92%.[1] This could lead to so-called *ex post selection* and the reporting of spurious heterogeneity that, in fact, resulted from so-called false positives.

The disadvantages of *ex post selection* of significant effects have been widely recognized in the programme-evaluation literature. For example, in randomized experiments researchers may be required to define their analysis plan for heterogeneity prior to the experiment to avoid only reporting (and searching for) significant effects (e.g., Casey, Glennerster, and Miguel, 2012, Olken, 2015). However, these pre-analysis plans are inflexible and usually not demanded (by funding bodies or editors of economic journals) in the common case of observational studies. An alternative approach that partly alleviates the *ex post selection* problem is to report effect heterogeneity for all possible groups. For large-scale investigations, an approach that takes account of all possible differences might lead to very small groups and

[1] Assuming independent test statistics as an extreme case ($1 - 0.95^{50} = 0.92$).

thus imprecise estimates. Further, the large number of different results makes it difficult to report the results in an intuitive way.

A developing part of the literature proposes to use machine learning algorithms (adapted for causal analysis) to systematically search for groups with heterogeneous effects (see, e.g., the review of Athey and Imbens, 2017b). Potentially, machine learning approaches are attractive because they could provide a principled approach to heterogeneity detection, which make it less likely to leave out important heterogeneities and can reduce concerns about the multiple testing problem. In addition, they enable flexible modelling and remain computationally feasible, even when the covariate space becomes high-dimensional and possibly exceeds the sample size.

In this study, we contribute to this newly developing literature in at least two ways. First, we systematically investigate effect heterogeneity of JSPs and report them in an interpretable way. We base the search algorithm for heterogeneity on many attributes of the unemployed persons as well as their caseworkers. For example, we consider the employment and welfare history of unemployed persons, socio-demographic characteristics, caseworkers' subjective employability ratings of their clients, and measures for the cooperativeness of caseworkers. The latter could uncover effect heterogeneity by different monitoring intensities, which we consider an important mechanism of JSPs (Behncke, Frölich, and Lechner, 2010a). Overall, we consider 1,268 different variables, including interactions and polynomials. Second, based on the detected heterogeneities, we document the potential of different assignment rules to improve JSPs' effects and cost-benefit efficiency.

Furthermore, we investigate the consistency of our findings across a variety of different machine learning algorithms. The (still young) causal machine learning literature is lacking large-scale sensitivity checks with regard to methodological choices in credible applications.

Obviously, the robustness of the results to possible misspecifications of the empirical model is essential for drawing coherent policy conclusions.

With respect to the active labour market programme (ALMP) evaluation literature that is based on informative data sets from administrative registers, it has become common practise to pursue a selection-on-observables strategy to identify the programme's effects (see, e.g., Imbens and Wooldridge, 2009, for standard econometric approaches and their properties, and e.g. Card, Kluve, and Weber, 2015, for an overview and a meta analysis of evaluation studies of active labour market programmes). We use exceptionally rich linked unemployed-caseworker data obtained from Swiss social security records.

For the investigation of effect heterogeneity, we combine Inverse Probability Weighting (IPW) with the so-called Modified Covariate Method (MCM) (Tian et al., 2014, Chen et al., 2017). The selection of relevant heterogeneity is carried out with Tibshirani's (1996) Least Absolute Shrinkage and Selection Operator (LASSO). For the quantification of the effects and their inference, we follow the sample splitting approach (see recent discussion in Rinaldo et al. 2016). We use half of the sample to select variables that are relevant to predict the size of the heterogeneous treatment effect, i.e. that are responsible for deviations from the average effects. We use the other half of the sample for inference on the (possibly low-dimensional) selected variables and the heterogeneous effects.

Our results suggest substantial effect heterogeneity of Swiss JSPs during the first six months after the start of participation. During this so-called 'lock-in' period, we observe negative effects for most participants. However, the size of the heterogeneity is strongly related to the characteristics of the unemployed. Consistent with the previous literature, participants with disadvantaged labour market characteristics benefit more from JSPs (e.g., Card, Kluve, and Weber, 2015). A major reason is that they face generally lower lock-in effects and, thus, these indirect programme costs are lower. Additionally, this study appears to be the first to uncover

substantial effect heterogeneity by residence status. We show that JSPs are more effective for foreigners, who have less access to informal job search networks compared to locals. For caseworker characteristics, however, there is only little heterogeneity. There is also no substantial effect heterogeneity beyond six months after the start of training. Finally, the paper presents easy-to-implement assignment rules which would improve the current assignment mechanism in a (almost) cost neutral way. An extensive sensitivity analysis shows that the main conclusions remain robust across a variety of different estimation methods.

In the next section, we provide information about the institutional background of the Swiss ALMP. In Section 3, we document the sample selection and show basic descriptive statistics. In Section 4, we discuss the econometric approach for a principled investigation of effect heterogeneity. In Section 5, we report the empirical findings and robustness checks. Section 6 explains our conclusions. Additional descriptive statistics, detailed information on the estimation of the selection procedures, and results for additional outcome variables, as well as extensive sensitivity analyses are reported in Online Appendices A-F.

2 Background

2.1 Swiss institutions

Switzerland is a federal country with 26 cantons and three major language regions (French, German, and Italian). It is a relatively wealthy country with approximately 78,000 CHF (approx. 77,000 US-Dollar) GDP per capita and a low unemployment rate of 3 to 4% (SECO, 2017, Federal Statistical Office, 2017). Unemployed persons have to register at the regional employment agency closest to their home.[2] The employment agency pays income maintenance. Benefits amount to 70 to 80% of the former salary depending on age, children, and past salary

[2] At the beginning of the unemployment spell, newly registered unemployed persons are often sent to a one-day workshop providing information about the unemployment law, obligations and rights, job search requirements, etc.

(see Behncke, Frölich, and Lechner, 2010b). The maximum benefit entitlement period is 24 months.

The yearly expenditures for Swiss ALMPs exceed 500 million CHF (Morlok et al., 2014). Unemployed persons can participate in a variety of different ALMPs. Gerfin and Lechner (2002) classify these ALMPs as (a) training courses, (b) employment programmes, and (c) temporary employment schemes. Training courses include job search, personality, language, computer, and vocational programmes. We focus on JSPs in this study, which is the most common ALMP in Switzerland (more than 50% of the assigned ALMPs are JSPs, see Huber, Lechner, and Mellace, 2017). JSPs provide training in effective job search and application strategies (e.g., training in résumé writing). Furthermore, actual applications are screened and monitored. JSPs are relatively short, with an average duration of about three weeks. Training takes place in class rooms. The employment agency covers the costs of training and travel. Participants are obliged to continue to search for jobs during the course.

In Switzerland, regional employment agencies have a large degree of autonomy, which is partly related to the country's federal organisation. Caseworkers make the decision to assign unemployed persons to a training course based on information about the unemployed person (e.g. employment history, subjective employability rating, etc.). Additionally, employment agency policies and federal eligibility rules are relevant for the assignment decision. The federal eligibility rules are rather vague. They imply, for example, that the training has to be necessary and adequate to improve the individual's employment chances. Caseworkers can essentially force the unemployed into such courses by threatening to impose sanctions. Unemployed persons have the option to apply to participate in such courses, but the final decision is always made by the caseworkers.

2.2 Related literature on job search programmes (JSPs)

An assignment to a JSP may affect the matching process and quality alignment between the participant and his or her potential new job (see, e.g., Blasco and Rosholm, 2011, Cottier et al., 2017). Push effects could occur if participants accept jobs with low matching quality because of actual or perceived sanctions or perceived future ALMP assignments. Push effects decrease the duration of unemployment, but may reduce employment stability. On the other hand, JSP participation could improve the visibility of suitable job vacancies and the efficiency of the application process, which may improve employment stability. Furthermore, many studies are concerned with the crowding-out of non-participants (see, e.g., Blundell et al., 2004, Crépon et al., 2013, Gautier et al., 2017).

Empirical evidence about the effectiveness of JSPs is mixed. The review studies of Card, Kluve, and Weber (2010, 2015) as well as Crépon and van den Berg (2016) document a weak tendency towards positive effects of JSPs, especially in the short-term.[3] However, for Swiss JSPs, the literature finds negative employment effects, which taper off one year after the start of participation (see Gerfin and Lechner, 2002, Lalive, van Ours, and Zweimüller, 2008). One reason for the ambiguous effectiveness of JSPs might be the different relative intensities of job search training and monitoring. Van den Berg and van der Klaauw (2006) are concerned that intensive monitoring reduces an informal job search, which might be a more efficient strategy than a formal job search for some unemployed persons. They suggest a formal job search is more effective for individuals with fewer labour market opportunities. Consistent with their arguments, Card, Kluve, and Weber (2015) document that JSPs are relatively more effective for disadvantaged participants. Vikström, Rosholm, and Svarer (2013) find slightly

[3] Meyer (1995) reports negative effects on unemployment benefit payments and positive earnings effects of JSPs in the US. Graversen and van Ours (2008) and Rosholm (2008) report positive effects of JSPs on the unemployment exit rate in Denmark. Wunsch and Lechner (2008) find JSPs have negative effects during the first two years after a programme begins, which fade out afterwards in Germany. They also show that training sequences are responsible for long lasting negative lock-in effects.

more positive effects of JSPs for women and younger participants. Dolton and O'Neill (2002) report negative employment effects of JSPs for men and insignificant effects for women five years after the programme begins. Surprisingly, the programme evaluation literature is lacking large-scale evidence about the effect heterogeneity of JSPs.

3 Data

3.1 General

The data we use includes all individuals who are registered as unemployed at a Swiss regional employment agency in the year 2003. The data contains rich information from different unemployment insurance databases (AVAM/ASAL) and social security records (AHV). This is the standard data used for many Swiss ALMP evaluations (e.g. Gerfin and Lechner, 2002, Lalive, van Ours, and Zweimüller, 2008, Lechner and Smith, 2007). We observe (among others) residence status, qualification, education, language skills, employment history, profession, job position, industry of last job, and desired occupation and industry. The administrative data is linked with regional labour market characteristics, such as the population size of municipalities and the cantonal unemployment rate. The availability of extensive caseworker information and their subjective assessment of the employability of their clients is what distinguish our data. Swiss caseworkers employed in the period of 2003 to 2004 were surveyed through a written questionnaire in December 2004 (see Behncke, Frölich, and Lechner, 2010a, 2010b). The questionnaire asked about the caseworker's aims and strategies and information about the regional employment agency.

3.2 Sample definition

In total, 238,902 persons registered as being unemployed in 2003. We only consider the first unemployment registration per individual in 2003. Each registered unemployed person is assigned to a caseworker. In most cases, the same caseworker is responsible for the entire unemployment duration of his or her client. If this is not the case, we focus on the first caseworker to avoid concerns about (rare) endogenous caseworker changes (see Behncke, Frölich, and Lechner, 2010b). We only consider unemployed persons aged between 24 and 55 years who receive unemployment insurance benefits. We omitted unemployed persons who apply for disability insurance benefits, persons whose responsible caseworker is not clearly defined, or persons whose caseworker did not answer the questionnaire (the response rate is 84%). We omitted unemployed foreigners with a residence permit that is valid for less than a year. Finally, we omitted unemployed persons from five regional employment agencies that are not comparable to the other regional employment agencies. This sample is identical to the data used in Huber, Lechner, and Mellace (2017). It contains 100,120 unemployed persons.

One concern regarding the treatment definition is the timing with respect to the elapsed unemployment duration prior to participation. Caseworkers may assign unemployed persons to job training programmes at essentially anytime during their unemployment spell. The dynamic or sequential programme assignment has received considerable attention in evaluation literature (see the discussions in Abbring and van den Berg, 2003, 2004, Fredriksson and Johansson, 2008, Heckman and Navarro, 2007, Lechner, 2009, Robins, 1986, Sianesi, 2004, among others). We consider a classical static evaluation model and define treatment as the first participation in a JSP during the first six months of unemployment (83% of JSP are assigned within the first six months of unemployment). We exclude individuals who participate in other ALMPs within the first six months of unemployment from the sample, such that our control group represents non-participants of all programmes (8,787 other ALMP participants are dropped). Potentially, this approach could lead to a higher share of individuals with better labour market

characteristics among the control group than among the training participants, because individuals in the control group may have possibly found another job prior to their potential treatment times. This would negatively bias the results. To overcome this concern, we randomly assign (pseudo) participation starts to each individual in the control group. Thereby, we recover the distribution of the elapsed unemployment duration at the time of training participation from the treatment group (similar to, e.g., Lechner, 1999, Lechner and Smith, 2007). To ensure comparability of the treatment definitions of the participants and non-participants, we only consider individuals who are unemployed at their (pseudo) treatment dates. This makes the groups of participants and non-participants comparable with respect to the duration of un-employment and ensures that the treated and control groups are eligible for programme participation at their respective assigned start dates.

The final sample contains 85,198 unemployed persons (Table A.1 in Online Appendix A provides the details of the sample selections steps). From this sample, 12,998 unemployed persons participate in a JSP and 72,200 are members of the control group. These 85,198 unemployed persons are assigned to 1,282 different caseworkers.

3.3 Descriptive statistics

Table 1 reports the means and standard deviations by JSP participation for some selected variables. During the first 6 months after training begins, JSP participants are fewer months employed than non-participants. The standardised difference is above 20.[4] During the first 12 and 31 months after training begins, JSP participants also have a shorter employment duration

[4] The standardised difference of variable X between samples A and B is defined as

$$SD = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{1/2\left(Var(\bar{X}_A) + Var(\bar{X}_B)\right)}} \cdot 100,$$

where $\bar{X}_A$ denotes the mean of sample A and $\bar{X}_B$ denotes the mean of sample B. Rosenbaum and Rubin (1983) consider a standardised difference of more than 20 as being 'large'.

than non-participants, but the standardised differences decline. During months 25 to 31 after training begins, the difference in the employment duration is minor.

Table 1: Descriptive statistics of some important variables by JSP participation status.

	Participants		Non-Participants		Std. Diff.
	Mean	S.D.	Mean	S.D.	
	(1)	(2)	(3)	(4)	(5)
Outcome: Months employed since programme start					
During first 6 months	1.21	1.93	1.94	2.44	23.29
During first 12 months	3.68	4.27	4.53	4.80	13.12
During first 31 months	15.30	12.49	15.59	12.85	1.60
During months 25 - 31	3.48	2.88	3.33	2.86	3.72
Characteristics of unemployed persons					
Female	0.45	-	0.44	-	0.58
Age (in 10 years)	3.73	0.88	3.66	0.86	5.59
Unskilled	0.22	-	0.23	-	1.80
Some qualification degree	0.60	-	0.56	-	5.19
Employability rating low	0.12	-	0.14	-	3.97
Employability rating medium	0.77	-	0.74	-	5.79
Employability rating high	0.11	-	0.12	-	3.62
# of unemp. spells in last 2 years	0.41	0.98	0.64	1.27	13.85
Fraction of months emp. in last 2 years	0.83	0.22	0.79	0.25	12.57
Past income (in 10,000 CHF)	4.58	2.02	4.16	2.05	14.50
Caseworker characteristics					
Female	0.45	-	0.41	-	6.94
Age (in years)	44.0	11.6	44.4	1.16	7.7
Tenure (in years)	5.54	3.23	5.86	3.31	6.84
Own unemp. experience	0.63	-	0.63	-	0.54
Vocational training degree	0.26	-	0.23	-	5.63
Local labour market characteristics					
German speaking REA	0.89	-	0.67	-	39.68
French speaking REA	0.08	-	0.25	-	33.30
Italian speaking REA	0.03	-	0.08	-	16.81
Cantonal unemployment rate (in %)	3.64	0.77	3.75	0.86	9.23
Cantonal GDP per capita (in 10,000 CHF)	5.13	0.92	4.92	0.93	15.75
# of caseworkers	989		1,282		
# of observations	12,998		72,200		

Note: We report unconditional means for all variables, standard deviations (S.D.) for all non-binary variables, and standardised differences between participants and non-participants. The descriptive statistics of all confounding variables used in this study are shown in Table B.1 of Online Appendix B. REA is the abbreviation for regional employment agency.

Furthermore, Table 1 documents descriptive statistics of the characteristics of the unemployed person, the characteristics of his or her caseworker, and local labour market

conditions. We report the descriptive statistics for additional control variables in Table B.1 of Online Appendix B. JSP participants have spent more months employed and received a higher income than non-participants in the last two years prior to the programme's start. We document minimal difference between the caseworkers of participants and non-participants.[5] JSP participants are more often registered at German-speaking regional employment agencies and live in cantons with better economic conditions (in terms of local GDP and unemployment rate) than non-participants.

4 Econometric approach

4.1 Parameters of interest

We describe the parameters of interest using Rubin's (1974) potential outcome framework. Following the conventional notation, we indicate random variables by capital letters and the realizations of these random variables by lowercase letters. The binary treatment dummy D_i indicates JSP participation. Let Y_i^1 denote the potential outcome (e.g., employment) when individual i ($i = 1, \dots, N$) participates in a JSP ($D_i = 1$). Conversely, Y_i^0 denotes the potential outcome when individual i is not participating in a JSP ($D_i = 0$). Obviously, each individual can either participate in a JSP or not, but both participation states cannot occur simultaneously. This implies only one potential outcome is observable. The observed outcome equals

$$Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i).$$

The causal effect of D on Y for individual i is

$$\gamma_i = Y_i^1 - Y_i^0.$$

[5] In Table B.1 of Online Appendix B we also show caseworker characteristics interacted with the language of the regional employment agency. For some interacted variables, we find strong differences between participants and non-participants.

However, we cannot identify the parameter γ_i without assumptions that are implausible in many applications (e.g., effect homogeneity). Nevertheless, group averages of γ_i may be identifiable under plausible assumptions. For example, the identification of the average treatment effect (ATE), $\gamma = E[\gamma_i]$, the average treatment effect on the treated (ATET), $\theta = E[\gamma_i|D_i = 1]$, and the average treatment effect on the non-treated (ATENT), $\rho = E[\gamma_i|D_i = 0]$, are standard econometric problems (see, e.g., Imbens and Wooldridge, 2009). Furthermore, conditional average treatment effects (CATEs) can potentially uncover effect heterogeneity based on exogenous pre-treatment variables Z_i chosen by the researcher based on the policy interest,

$$\gamma(z) = E[\gamma_i|Z_i = z] = E[Y_i^1 - Y_i^0|Z_i = z].$$

Knowledge about CATEs could help, e.g., to improve the assignment mechanism to JSPs.[6]

4.2 Econometric background and intuition

Machine learning methods are powerful tools for out-of-sample predictions of observable variables. However, the fundamental problem of causal analyses is the inability to observe individual causal effects because at least one potential outcome is unobservable. Recently, several methods have been proposed that apply machine learning methods in ways that overcome this fundamental problem (see, e.g., the reviews by Belloni, Chernozhukov, and Hansen, 2014, Horowitz, 2015, and Varian, 2014).

Concerning effect heterogeneity, Imai and Ratkovic (2013) suggest a LASSO-type algorithm while Athey and Imbens (2016) propose a regression tree method. Foster, Taylor, and Ruberg (2011) apply random forest algorithms to estimate effect heterogeneity. These algorithms are flexible and are effective at capturing multi-dimensional and non-linear

[6] Additional parameters are CATEs for JSP participants $\theta(z) = E[\gamma_i|D_i = 1, Z_i = z]$ and CATEs for non-participants $\rho(z) = E[\gamma_i|D_i = 0, Z_i = z]$. The parameters $\gamma(z)$, $\theta(z)$, and $\rho(z)$ can differ from each other when Z_i differs from X_i (which is the case in our application). However, we are interested in the heterogeneities for a random unemployed person with specific characteristics because this mirrors the decision problem of the caseworker. Thus, we focus on $\gamma(z)$.

interactions among covariates. Imai and Strauss (2011), Green and Kern (2012), and Taddy et al. (2015) propose alternative Bayesian machine learning methods to estimate effect heterogeneity. Grimmer, Messing, and Westwood (2016) do not attempt to use the best method available. Instead, they suggest combining many different machine-learning tools to estimate the conditional treatment responses. Athey and Wager (2017), Qian and Murphy (2011), Xu et al. (2015), and Zhao et al. (2012) focus on the estimation of individualized treatment rules, which primarily focus on decision rules instead of effect heterogeneity.[7]

All of these studies consider heterogeneity in randomized experiments. In many fields of economics, randomized experiments are expensive and minimally socially acceptable. Therefore, we consider a selection-on-observables identification strategy (e.g. Imbens, Wooldridge, 2009). A promising approach to estimate group specific causal effects in non-experimental approaches is the Modified Covariate Method (MCM).[8]

To gain some intuition about the MCM, we first consider that participation in a programme is randomly assigned to 50% of the unemployed persons. Accordingly, in this introductory example there is no need to adjust for selection into training participation. Throughout the analyses the first element in Z_i is a constant term ($Z_{i0} = 1$) and the remaining elements of Z_i contain additional $p \geq 1$ pre-treatment variables that are potentially related to the effect heterogeneity in which the researcher is interested. A standard approach to estimate effect heterogeneity is to use the interaction model,

$$Y_i = Z_i\beta_s + D_i Z_i \delta + u_i. \quad (1)$$

[7] Closely related is the study of Ciarleglio et al. (2015), who propose a method to select the optimal treatment conditional on observed individual characteristics. Zhao et al. (2015) investigate the optimal dynamic order of sequential treatments.

[8] Furthermore, Zhang et al. (2012) develop alternative non-experimental approaches for a principled effected heterogeneity search, which is an adaptation of the Modified Outcome Method (MOM) (Signorovitch, 2007). We describe the MOM in Online Appendix F.1. For one of the robustness checks, we replicate our results using the MOM. Furthermore, the tree and forest methods of Athey and Imbens (2016) and Wager and Athey (2017) are applicable in non-experimental settings. All robustness checks are provided in Section 5.7 and Online Appendix F. The main findings are not altered.

The first term on the right side of equation (1) provides a linear approximation of the conditional expectation of the potential outcome under non-participation, $E[Y_i^0|Z_i = z] = z\beta_s$. We call this the main effects in the following. The second term on the right-hand side of equation (1) provides a linear approximation of the CATE:

$$\gamma(z) = z\delta = E[Y_i^1 - Y_i^0|Z_i = z].$$

Vansteelandt et al. (2008) point at possible sensitivities of the empirical model in equation (1) when the main effects are miss-specified. Tian et al. (2014) propose to transform the treatment dummy $T_i = 2D_i - 1$ and rearrange the interaction model in equation (1) to:

$$Y_i = Z_i\beta_t + \frac{T_i Z_i}{2}\delta + v_i. \quad (2)$$

The treatment indicator shifts from $D_i \in \{0,1\}$ to $T_i/2 \in \{-0.5,0.5\}$. The modification does not alter the coefficient vector δ. However, this transformation alters the main effects. In equation (2), $E[Y_i|Z_i = z] = z\beta_t$ is the linear approximation of the conditional expectation of the observed outcome. Notice that $Cov(Z_{ij}, T_i Z_{ik}) = Cov(Z_{ij}, Z_{ik})E[T_i] = 0$ for $j, k \in \{1, \dots, p\}$. The first equality holds under random assignment of training participation and the second equality holds because $E[T_i] = 0$.[9] Accordingly, the right hand terms of equation (2) are independent of each other and we can estimate the coefficients β_t and δ in two separate regressions. For example, we can estimate CATEs with the model

$$Y_i = \frac{T_i Z_i}{2}\delta + \varepsilon_i,$$

which is the baseline model of the MCM. The MCM is suitable when only the interaction effects and not the main effects are of interest. Parsimony and robustness to misspecification of the main effects are two advantages of the MCM compared to the specification in equation (1). We can adopt the basic idea of the MCM to non-experimental identification strategies (see Chen et

[9] In contrast, $Cov(Z_{ij}, D_i Z_{ik}) = Cov(Z_{ij}, Z_{ik})E[D_i] = Cov(Z_{ij}, Z_{ik})/2$, which can be different from zero.

al., 2017). Furthermore, we can combine the MCM with different machine learning methods to select the variables for heterogeneity. Procedure 1 summarises our (main) estimation algorithm of the adapted MCM approach, which we describe in detail below.

4.3 Identification

In addition to the pre-treatment variables included in the vector Z_i (which are potentially related to effect heterogeneity), we consider the possibility of confounding variables, which are included in the vector X_i. Confounders are pre-treatment variables that jointly affect the probability to participate in a JSP and the employment outcome. The vector Z_i may be larger, smaller, partially, or fully overlapping with X_i depending on the question under investigation.

Assumption 1 (Conditional independence): $Y_i^1, Y_i^0 \coprod \mathrm{D}_i | \mathrm{X}_i = \mathrm{x}, \mathrm{Z}_i = \mathrm{z}$ for all values of x and z in the support of X and Z.

Assumption 2 (Common support): $0 < P(D_i = 1 | X_i = x, Z_i = z) = p(X_i, Z_i) < 1$ for all values of x and z in the support (of interest) of X and Z.

Assumption 3 (Exogeneity of controls): $X_i^1 = X_i^0$ and $Z_i^1 = Z_i^0$.

Assumption 4 (Stable Unit Treatment Value Assumption, SUTVA): $Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i)$.

Assumption 1 states that the potential outcomes are independent of programme participation conditional on the confounding pre-treatment variables. The plausibility of this assumption is justified by the availability of a detailed set of confounding variables containing characteristics of the unemployed and the caseworkers. The studies of Biewen et al. (2014) and Lechner and Wunsch (2013) discuss the selection of confounders in ALMP evaluations based on rich administrative data. Within the employment agency, caseworkers have high autonomy to decide about assignment of JSPs. Our data contain the same objective measures about labour market history, education and socio-demographics of the unemployed, as well as local labour market characteristics that are observable to the caseworkers when choosing who participates

in JSPs. We observe caseworkers' subjective ratings of the employability of their clients. Furthermore, we observe detailed information about the caseworkers' characteristics and counselling styles. These are potential confounders, because caseworker characteristics might affect the probability of JSP participation and labour market outcomes simultaneously.

According to Assumption 2, the conditional probability to participate in a JSP is bounded away from zero and one. The common support assumption has to hold when conditioning jointly on X and Z. We enforce common support by trimming observations below the 0.5 quantile of participants and above the 99.5 quantile of non-participants.[10] This procedure shows good finite sample performance in the study Lechner and Strittmatter (2017). Assumption 3 requires exogeneity of confounding and heterogeneity variables. To account for this assumption, we only use control variables that are determined prior to the start of JSP participation. Assumption 4 excludes spillover effects between participants and non-participants.

Theorem 1 (Identification): Under Assumptions 1-4 (and regularity conditions ensuring the existence of appropriate moments) the following equality holds:

$$\gamma(z) = E_{X|Z=z}[E(Y_i|\, D_i = 1, X_i = x, Z_i = z)|Z_i = z] \\ - E_{X|Z=z}[E(Y_i|\, D_i = 0, X_i = x, Z_i = z)|Z_i = z].$$

Thus $\gamma(z)$ are identified from observable data on $\{Y_i, D_i, Z_i, X_i\}_{i=1}^{N}$. For completeness, the proof of Theorem 1 is in Online Appendix C (see also, e.g., Rosenbaum and Rubin, 1983).

4.4 Search for effect heterogeneity

Chen et al. (2017) outline how we can combine MCM with Inverse Probability Weighting (IPW), a standard approach to balance covariates in observational studies (see, e.g., Hirano,

[10] In total, we trim 6,767 observations (579 participants, 6,188 non-participants).

Imbens, and Ridder, 2003, Horvitz and Thompson, 1952). We can estimate the parameter vector δ using Weighted Ordinary Least Squares (WOLS), i.e. by minimising the objective function

$$\underset{\hat{\delta}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \widehat{w}(D_i, X_i, Z_i) T_i \left(Y_i - \frac{T_i Z_i}{2} \hat{\delta} \right)^2 \right], \tag{3}$$

with the IPW weights

$$\widehat{w}(D_i, X_i, Z_i) = \frac{\frac{D_i - \hat{p}(X_i, Z_i)}{\hat{p}(X_i, Z_i)(1 - \hat{p}(X_i, Z_i))}}{D_i \sum_{i=1}^{N} \frac{D_i}{\hat{p}(X_i, Z_i)} + (1 - D_i) \sum_{i=1}^{N} \frac{1 - D_i}{1 - \hat{p}(X_i, Z_i)}},$$

which we calculate using the estimated propensity score $\hat{p}(X_i, Z_i)$. In our baseline model, we adapt the propensity score specification of Huber, Lechner, and Mellace (2017), which we report in Table D.1 of Online Appendix D. The denominator of the IPW weights causes a small sample adjustment (see, e.g., Busso, DiNardo, and McCrary, 2014). In equation (3), we multiply the IPW weights by T_i, such that the weights are positive.

The variables included in Z_i, which are potentially related to effect heterogeneity, consist of individual and caseworker characteristics, their second order interactions, up to fourth order polynomials, and logarithms of non-binary variables. Additionally, we consider dummy variables for the 103 employment agencies as well as 29 category dummies for previous industry and 29 category dummies describing the previous job. In total, this leads to 1,268 heterogeneity variables that we consider in the analyses.[11]

In our main specifications, we employ LASSO estimators. The weighted LASSO estimator of the MCM minimizes the objective function,

$$\underset{\hat{\delta}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \widehat{w}(D_i, X_i, Z_i) T_i \left(Y_i - \frac{T_i Z_i}{2} \hat{\delta} \right)^2 \right] + \lambda \sum_{j=1}^{p} \left| \hat{\delta}_j \right|, \tag{4}$$

[11] We exclude binary variables where less than 1% of (non-) participants show values of 0 or 1. Furthermore, we keep only one variable of variable combinations that show correlations of larger magnitude than ±0.99 to speed up computation.

where we add a penalty term for the sum of the absolute values of the coefficients of the p variables appearing in Z. Importantly, we do not penalize the constant $\hat{\delta}_0$. The penalising parameter λ specifies the amount of penalisation. If $\lambda = 0$, then equation (4) is equivalent to the WOLS model in equation (3). However, when $\lambda > 0$ some coefficients are shrunken towards zero. For sufficiently large values of λ, some (or all) coefficients are exactly zero. Therefore, the LASSO serves as a model selector, omitting variables with little predictive power from the model.[12] A challenge is the optimization of the penalty term, such that only the relevant predictors of the effect heterogeneity remain in the model. Too low penalties lead to overfitting, too high penalties lead to models that miss important variables (i.e., we have a bias-variance trade-off).

We apply 10-fold cross-validation to find the penalty term λ with the best out-of-sample performance in terms of mean-squared-error (MSE) (e.g., Bühlmann and van de Geer, 2011).[13] The LASSO coefficients are biased when $\lambda > 0$ (regularisation bias, see, e.g., Zou, 2006). For this reason, we use the so-called Post-LASSO estimates to calculate the MSE. We obtain the Post-LASSO coefficients from a WOLS model, which includes all variables with non-zero coefficients in the respective LASSO model (see, e.g. Belloni, Chernozhukov, Hansen, 2013). We choose the LASSO model with the penalty parameter λ that minimises the Post-LASSO MSE.[14]

There is no need to specify the main effects in the MCM approach. Nevertheless, Tian et al. (2014) and Chen et al. (2017) show that accounting for the main effects can improve the finite sample performance of the MCM because they can absorb variation in the outcome, which

[12] The larger the values of λ the fewer variables remain in the model. By gradually increasing the penalty term one can obtain a path from a full model to a model that only contains the parameter $\hat{\delta}_0$.

[13] Chetverikov, Liao, and Chernozhukov (2017) discuss the properties of K-fold cross-validation in the context of LASSO. They derive bounds for the prediction errors of cross-validated LASSO estimators.

[14] In robustness checks, we base the selection of the penalty parameter on the LASSO MSE. The main results are not altered.

is unrelated to the effect heterogeneity. In Online Appendix F.2, we document two ways to implement an efficiency augmenting procedure.

Note that in case Z contains additional variables to the confounders X, there is some concern that including Z in the estimation of the propensity score might inflate the propensity score without removing additional selection bias. Therefore, our main specification is based on $p(x)$ only. We also estimate specifications allowing Z to enter the propensity score as well, $p(x,z)$ (see Appendix F.5). However, besides decreasing the precision of the estimates, the main results are not altered.

4.5 Estimation of CATEs

To avoid the situation in which the LASSO approach models idiosyncratic within-sample effects, we randomly partition the sample into two equal sized parts. We assume independence between the two samples. We use the first sample to select the relevant effect heterogeneity variables (training sample). We use the second sample for the estimation of a WOLS model including all selected heterogeneity variables (estimation sample). This is called the 'honest' inference procedure (see the discussion about the general properties, e.g., in Fithian, Sun, and Taylor, 2017).

The CATE for individual i is estimated as $\hat{\gamma}(Z_i) = Z_i\hat{\delta}$. All coefficients of variables not selected in the training sample are set to zero. The coefficients of the selected variables are estimated in the estimation sample and extrapolated to the full sample. The medical and biometric literature calls $\hat{\gamma}(Z_i)$ individualised treatment effects (ITE) (e.g., Chen et al., 2017). The estimates of $\hat{\delta}$ vary with respect to the random sample split. To reduce the dependency of the results on a particular split, we run the analyses $S = 30$ times with different random splits. We calculate the individualised CATEs, $\hat{\gamma}_s(Z_i) = Z_i\hat{\delta}_s$, for each split, where the Post-LASSO coefficients, $\hat{\delta}_s$, are from the random sample split s. We use these parameters to calculate the

aggregated CATEs, $\bar{\gamma}(Z_i) = \frac{1}{S}\sum_{s=1}^{S}\hat{\gamma}_s(Z_i)$. This procedure is in the spirit of bootstrap aggregation (`bagging`) in machine learning literature (see, e.g., Breiman, 1996). It reduces model dependency and smooths the estimated CATEs, but the estimation model of $\bar{\gamma}(Z_i)$ is more difficult to interpret than the model of $\hat{\gamma}(Z_i)$. To understand which factors influence the aggregated CATEs, we report averages by different groups,

$$\bar{\gamma}_g = \frac{1}{\sum_{i=1}^{N} G_i}\sum_{i=1}^{N} G_i\,\bar{\gamma}(Z_i),$$

where the binary variable G_i indicates whether individual i belongs to the group ($G_i = 1$) or not ($G_i = 0$). These groups could, for example, be all JSP participants, all non-participants, or unemployed persons with specific characteristics.

4.6 Variance estimation

It appears natural to estimate the variance with a bootstrap approach over the whole estimation algorithm, including the variable selection step. However, this is computationally infeasible for a reasonable number of bootstrap replications. Thus, we use a computationally feasible bootstrap approach in which we fix the selected heterogeneity variables in each sample split.

First, we draw a random bootstrap sample b (with replacement) clustered on the caseworker level. Second, for each sample split, we align the observations in the bootstrap sample to the observations in the original estimation sample. We only keep observations that we observe in both the bootstrap and the estimation sample. Third, based on these samples, we re-estimate the CATEs for each sample split using the heterogeneity variables selected in the original training sample of the respective sample split. We repeat these three steps 1,000 times. This procedure takes into account the dependencies that stem from overlapping observations across sample splits.

Procedure 1: Estimation algorithm of the adapted MCM.

Step 1	Estimate propensity score $\hat{p}(X_i, Z_i)$ and calculate the IPW weights.
Step 2	a) Randomly split the sample into training and estimation sample s. b) Select the relevant heterogeneity variables in the training sample using the LASSO approach with or without efficiency augmentation (explained in Appendix F.2). c) Estimate the coefficients $\hat{\delta}_s$: (i) Set the coefficients of deselected variables to zero. (ii) Estimate the coefficients of the selected variables in the estimation sample. d) Calculate $\hat{\gamma}_s(Z_i) = Z_i\hat{\delta}_s$ for the full sample.
Step 3	a) Repeat Step 2 *S* times. b) Calculate the aggregated CATEs $\bar{\gamma}(Z_i) = \frac{1}{S}\sum_{s=1}^{S}\hat{\gamma}_s(Z_i)$ and group averages of CATEs $\bar{\gamma}_g = \frac{1}{\sum_{i=1}^{N} G_i}\sum_{i=1}^{N} G_i\,\bar{\gamma}(Z_i)$.
Step 4	Bootstrap the variance of $\bar{\gamma}(Z_i)$ and $\bar{\gamma}_g$. (For computational feasibility, we do not re-estimate Step 2b) in the bootstrap replications.)

For each sample split s and bootstrap replication b we obtain the bootstrapped CATEs, $\hat{\gamma}_{sb}(Z_i) = Z_i\hat{\delta}_{sb}$. The aggregated bootstrapped CATEs are $\bar{\gamma}_b(Z_i) = \frac{1}{S}\sum_{s=1}^{S}\hat{\gamma}_{sb}(Z_i)$. We estimate the standard error for the aggregated CATEs with

$$\hat{\sigma}_i = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{\gamma}_b(Z_i) - \frac{1}{B}\sum_{b=1}^{B}\bar{\gamma}_b(Z_i)\right)^2},$$

and the standard errors of CATEs by groups with

$$\hat{\sigma}_g = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{\gamma}_{gb} - \frac{1}{B}\sum_{b=1}^{B}\bar{\gamma}_{gb}\right)^2},$$

where $\bar{\gamma}_{gb}$ is the estimate of $\bar{\gamma}_g$ in the respective bootstrap replication b.

5 Results

5.1 Propensity score model

Table D.1 in Appendix D reports the average marginal effects of the estimated propensity score model. The propensity score estimates serve as inputs into the matching algorithm. The results

confirm the impression from the descriptive statistics in Table 1, namely that the participation probability is generally increasing with previous labour market success. Unemployed persons with good labour market opportunities have a greater probability to participate in a JSP. Such a selection of training participants is called 'cream-skimming' (e.g., Bell and Orr, 2002). The effect of training is not necessarily higher for participants with good labour market opportunities, because these participants would have good labour market opportunities even in the absence of training (see, e.g., discussion in Berger, Black, and Smith, 2000).

When performing matching, it is a best practice to check for potential issues of (i) insufficient support in the propensity scores across treatment states that may result in incomparable matches as well as large matching weights of some non-treated observations with specific propensity scores; and (ii) imbalances in covariates after matching (due to inappropriate propensity score specifications). We document the distribution of the baseline propensity score in Figure D.1 of Online Appendix D. Furthermore, we document the balancing of the control variables after matching in Table D.2 of Online Appendix D. We find only small imbalances between JSP participants and non-participants. The standardised differences are always below three.

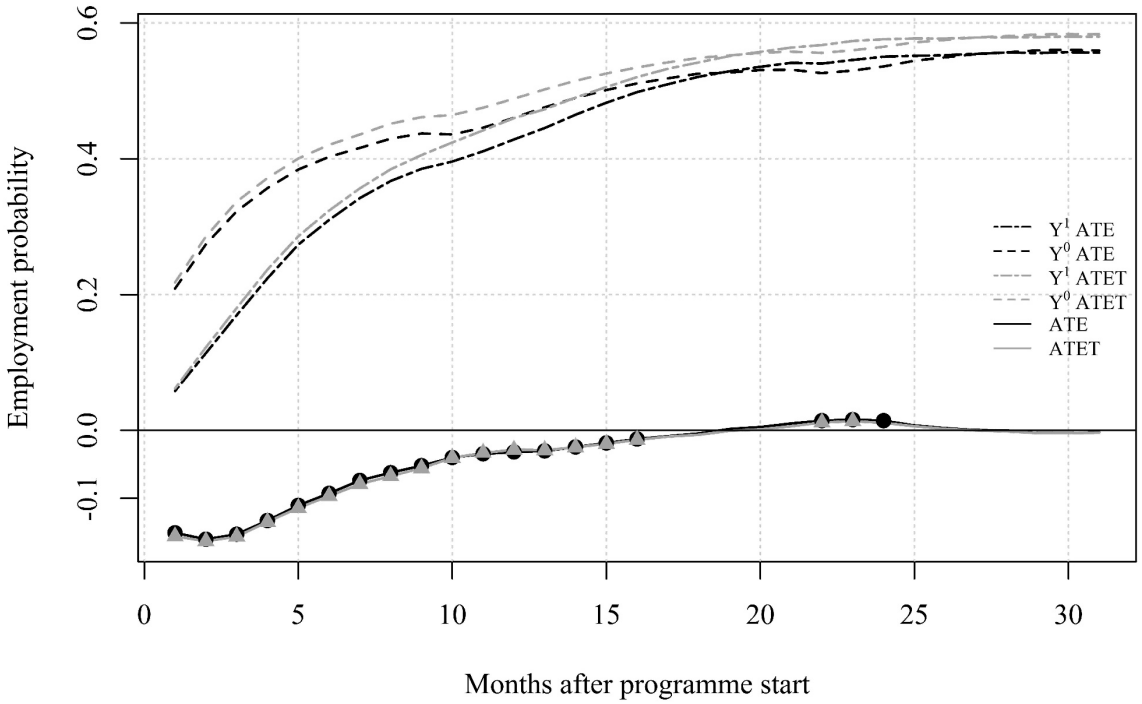
5.2 Average effects

Figure 1 shows the estimated potential outcomes and average programme effects on employment for each of the first 31 months after the programme's start. We observe substantial negative lock-in effects. The employment probability in the first three months is about 15 percentage points lower for JSP participants compared to non-participants. However, differences in the two groups' employment probability disappear after 16 months. In months 22 to 24 after a programme's start, we find small positive effects. But this seems to be only of short duration. Overall, the long-term effects are insignificant and close to zero. The negative lock-in effects are consistent with the findings of the previous Swiss JSP evaluations (e.g.,

Gerfin and Lechner, 2002, Lalive, van Ours, and Zweimüller, 2008). Moreover, the effectiveness of JSPs is also negative in other countries (see e.g., Dolton and O'Neil, 2002, Wunsch and Lechner, 2008). It is possible that participants reduce the intensity of informal job search during participation in a JSP, which could explain negative employment effects.

Searching for effect heterogeneity in each month after a programme's start is computationally expensive and hard to intuitively summarise (at least if it varies over time). Therefore, we estimate the effects of JSP participation on cumulated months employed during the first 6, 12, and 31 months after a programme begins, as well as during months 25 to 31. Table 2 shows the respective average effects that mirror the findings in Figure 1. The lower employment probabilities after programme participation translate into an average decline of 0.8 employment months (≈ -24 days) during the first six months after the start of participation. This decreases to -1.1 months (≈ -33 days) during the first 12 and 31 months. We find no significant employment effects during the months 25 to 31 after the start of participation.

Figure 1: ATE, ATET, and potential outcome levels by months since the start of JSP participation.



Note: We estimate the ATE and ATET separately for each of the first 31 months after start of JSP participation. Circles/triangles indicate significant effects at the 5% level. We obtain standard errors from a clustered bootstrap at caseworker level with 4,999 replications.

Table 2: ATE, ATET, and ATENT by duration since the start of JSP participation.

Months employed since start of participation	ATE		ATET		ATENT	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
	(1)		(2)		(3)	
During first 6 months	-0.80***	(0.02)	-0.82***	(0.02)	-0.80***	(0.02)
During first 12 months	-1.10***	(0.05)	-1.13***	(0.04)	-1.09***	(0.05)
During first 31 months	-1.14***	(0.14)	-1.20***	(0.13)	-1.12***	(0.15)
During months 25-31	-0.007	(0.03)	-0.011	(0.03)	-0.007	(0.04)

Note: We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

5.3 Effect heterogeneity

Table 3 reports the estimated heterogeneity coefficients, $\hat{\delta}$, obtained from one of the considered random partitions into training and estimation samples.[15] The coefficients are the marginal effects of the respective variables on the treatment effect of JSP (as opposed to the marginal effects of the respective variables on the outcome level in standard linear regression models).

The first column of Table 3 reports the estimated coefficients for the outcome cumulated employment during the first six months after training participation begins. In this specification, the Post-LASSO estimation selects 17 out of 1,268 potential variables. In the estimation sample, five of these variables are significant; for example, the treatment effect increments by 0.3 months ($\approx$ 9 days) for unskilled workers with previous earnings below 25,000 CHF a year (see row 3). When all other selected variables equal zero, the predicted effect of JSP participation for unskilled workers with previous earnings below 25,000 CHF a year would be $-0.89 + 0.3 = -0.59$ months employment ($\approx$ -17 days). However, we must be cautious when interpreting the model, because it is selected to maximise prediction power, which might differ from the structural (causal) model (see, e.g., discussion in Mullainathan and Spiess, 2017).

[15] We omit the coefficients of the main effects because they are only used for the efficiency augmentation and irrelevant for the interpretation.

Table 3: Post-LASSO coefficients for selected outcome variables.

	Months employed during first 6 months after the start of participation		Months employed during first 12 months after the start of participation	
	Coef.	S.E.	Coef.	S.E.
	(1)		(2)	
Constant	-0.89***	(0.05)	-1.29***	(0.09)
# of unemp. spells in last two years	0.06	(0.12)	-	-
Unskilled × past income 0 - 25k	0.30***	(0.11)	0.53	(0.53)
Skilled w/o degree × same gender like CW	0.20	(0.21)	-	-
Skilled w/o degree × age difference between unemployed & CW	-0.01	(0.01)	-	-
# of unemp. spells in last 2 years × age of CW	0.00	(0.00)	-	-
# of unemp. spells in last 2 years × medium city size	-0.05	(0.06)	-0.13	(0.14)
# of unemp. spells in last 2 years × past income 0 - 25k	-0.04	(0.06)	-0.10	(0.14)
# of unemp. spells in last 2 years × prev. job unskilled	0.04	(0.05)	0.21*	(0.13)
# of unemp. spells in last 2 years × same gender like CW	-0.01	(0.05)	-	-
CW has own unemp. experience × prev. job unskilled	0.19**	(0.09)	0.34*	(0.21)
Foreigner with perm. residence permit × past income 25 - 50k	0.19	(0.12)	-	-
Small city × past income 50 - 75k	-0.16*	(0.09)	-0.26	(0.20)
Single household × no emp. spell last 2 years	-0.17**	(0.08)	-	-
Single household × prev. job unskilled	0.16	(0.11)	-	-
Prev. job primary sector × age difference between unemp. person & CW	-0.02**	(0.01)	-	-
Prev. job restaurant	-0.01	(0.12)	-	-
Prev. job tourist sector	-0.09	(0.12)	-	-
Unskilled × prev. job unskilled	-	-	-0.22	(0.64)
# of unemp. spells in last 2 years × unempl. & CW have primary education	-	-	0.19**	(0.08)
CW has vocational training degree × past income 50 - 75k	-	-	-0.13	(0.30)
Past income 25 - 50k × unskilled	-	-	0.14	(0.24)
# of emp. spells past 5 years × prev. job in primary sector	-	-	-0.24	(2.16)
Prev. job in primary sector × unskilled	-	-	-0.19	(0.53)
Regional emp. agency No. 44	-	-	-0.68	(0.52)
# of selected variables	17 of 1,268		13 of 1,268	

Note: We apply one-step efficiency augmentation. We partition the data randomly into selection and estimation sample. We choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. We report results for additional outcomes in Table E.1 of Online Appendix E. CW is the abbreviation for caseworker. 25 - 50k is the abbreviation for 25,000-50,000 CHF. 50 – 75k is the abbreviation for 50,000-75,000 CHF.

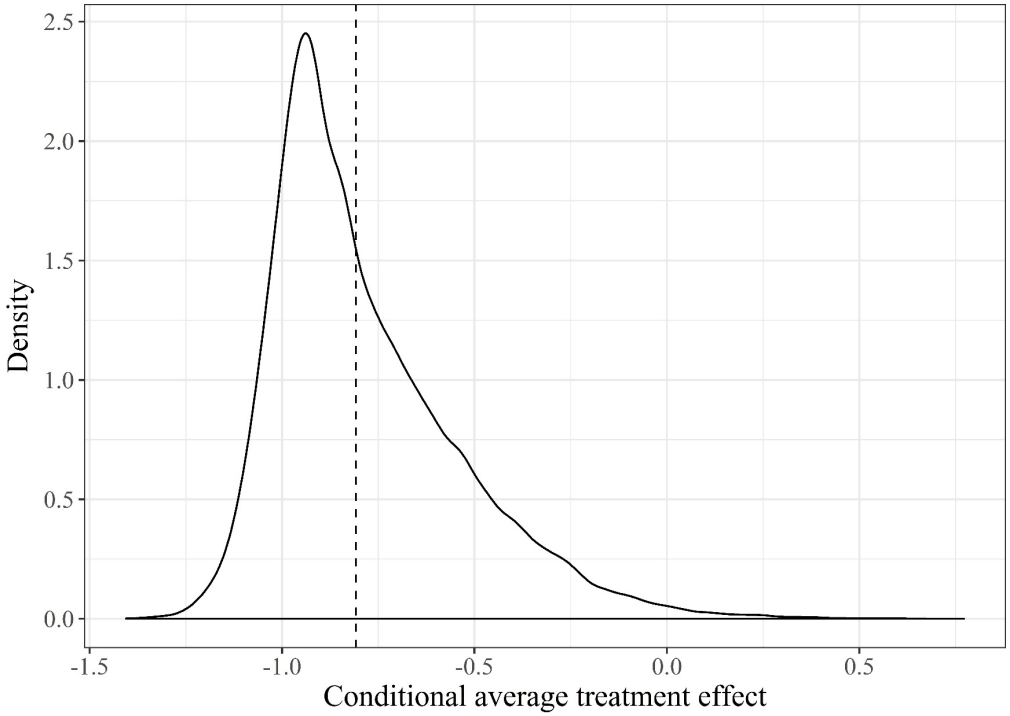
The second column of Table 3 shows the coefficients for the thirteen selected heterogeneity variables for the outcome cumulated employment during the first twelve months after training participation begins. The selected heterogeneity variables are partially overlapping between the two outcomes (comp. column 1 and 2). In Table E.1 in Online Appendix E, we report the selected heterogeneity parameters for the outcome cumulated

months employed in the first 31 months after training participation begins. We omit the results for the outcome cumulated months employed between months 25 to 31 after training participation begins, because we do not detect any effect heterogeneity in the considered sample split.

To improve precision and check the sensitivity of our results, we investigate the Post-LASSO models for different random sample splits. For each random partition, we obtain different Post-LASSO models (Table F.6 in Online Appendix F documents the number of selected variables in the different random sample splits). This is unsurprising, because many of the variables we consider are highly correlated (e.g., different measures of the employment history). Therefore, the same CATE can be obtained from different Post-LASSO models, each considering different variables or different functions of variables. Table F.1 in Online Appendix F documents the average correlation between CATEs for different sample splits. The correlations are positive and relatively large. Accordingly, the CATEs are highly consistent across the considered sample splits. This confirms that the selected models are not identical, but each model essentially predicts the similar CATEs.

One approach to get an overview of the detected heterogeneities is to plot the distribution of the predicted effects. Therefore, Figure 2 reports the distribution of the aggregated CATEs of JSPs on cumulated months employed during the first six months after participation begins. The figure documents substantial variation in the aggregated CATEs. For most groups of unemployed persons the aggregated CATE of JSP participation is between -0.8 and -1 months of employment (approximately a decline of between 24 and 30 days). However, the CATEs are less negative or even positive for a non-negligible fraction of the unemployed persons. This points at potential ways to improve assignment to a JSP.

Figure 2: Distribution of aggregated CATEs for months employed during first six months after the start of participation.



Note: Kernel smoothed distribution of average predicted individual effects. Gaussian kernel with bandwidth 0.02, chosen by Silverman's rule-of-thumb. We apply one-step efficiency augmentation. We partition the data randomly into selection and estimation sample. We choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. The dashed vertical line shows the ATE.

Table 4 reports summary statistics for the aggregated CATEs. For all outcomes, the means of the aggregated CATEs are close to the (semi-parametrically) estimated ATEs (comp. Table 2). This confirms that the estimation of the aggregated CATEs works well, on average. For all outcomes, the median is slightly lower than the mean. This suggests a right-skewed distribution (similar to Figure 2). We find substantial heterogeneity for the outcomes cumulated months employed during the first 12 months and the first 31 months after the start of JSP participation. After 12 months, the JSP effect ranges from minus two to plus two employment months. After 31 months, the JSP effect ranges from minus three to plus three employment months. However, for the outcome cumulated months employed between month 25 and 31 after the start of JSP participation we find little heterogeneity.

Table 4: Descriptive statistics of aggregated CATEs.

Months employed since start of participation	Mean	Median	S.D.	Min.	Max.	Mean S.E.
	(1)	(2)	(3)	(4)	(5)	(6)
During first 6 months	-0.78	-0.84	0.25	-1.41	0.77	0.07
During first 12 months	-1.10	-1.20	0.32	-2.09	1.44	0.10
During first 31 months	-1.13	-1.25	0.60	-3.79	4.12	0.23
During months 25-31	-0.04	-0.05	0.06	-0.32	0.48	0.04

Note: We obtain CATEs from aggregating CATEs from 30 different random sample splits. Standard deviations are abbreviated with S.D. in column (3). Column (6) shows mean standard errors of CATEs.

Accordingly, the MCM successfully discovers substantial effect heterogeneity. However, interpretation of the results is not easily accessible, because the underlying functions are too complex. Figure 2 and Table 4 document two ways to aggregate the results. However, we want to go beyond these abstract descriptions and make explicit policy recommendations. In the next section, we marginalise the effects for specific variables of interest. This enables us to reveal more of the CATEs' structure. Afterwards, we focus on the implementation of specific JSP assignment rules.

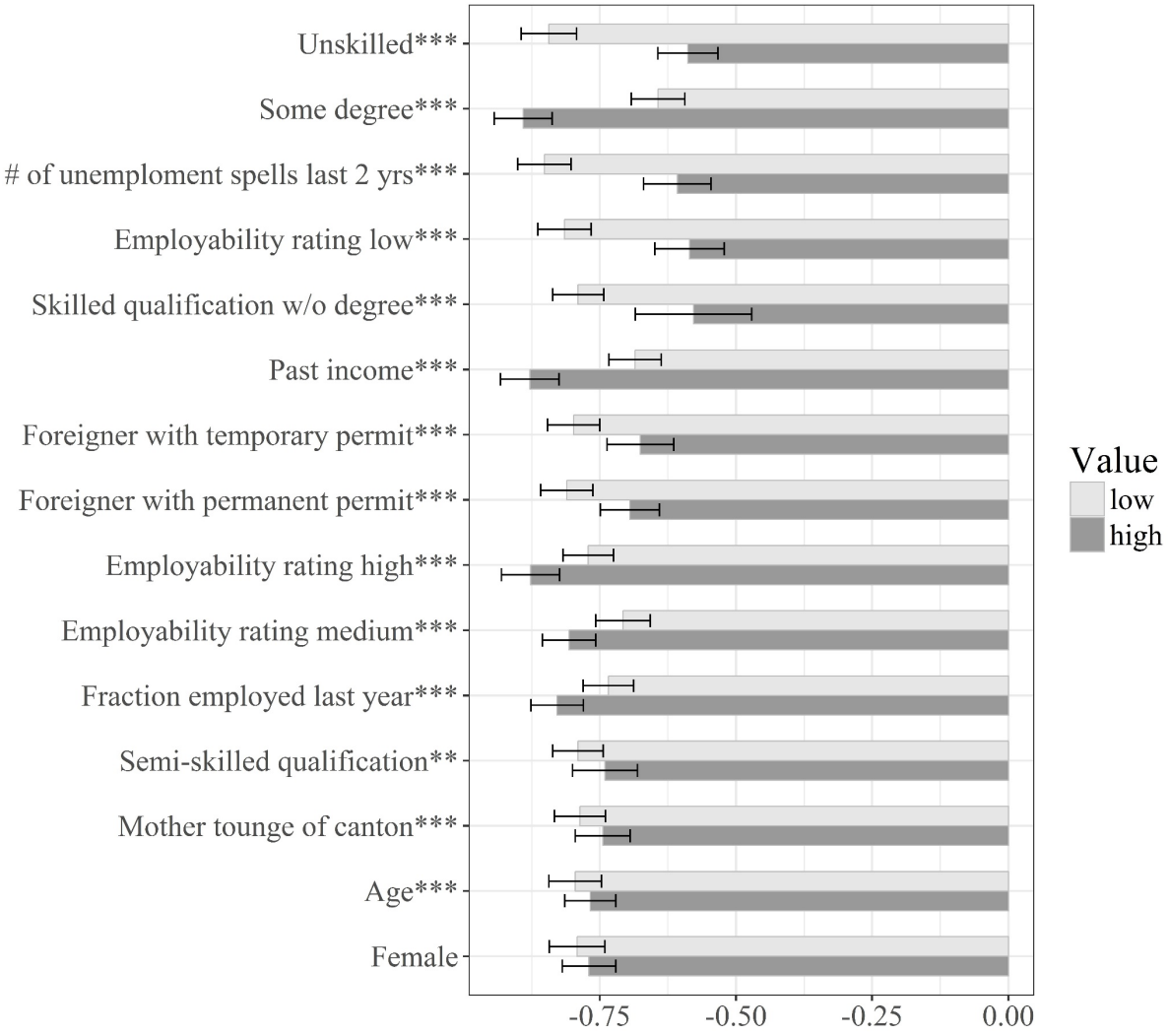
5.4 Effect heterogeneity by selected variables

In this section, we average CATEs by characteristics of unemployed persons and their case-workers. For each characteristic, we partition the sample in two mutually exclusive groups (high $g = 1$ and low $g = 0$ group), by using a binary characteristic itself as indicator or by discretising at the median of non-binary characteristics. The parameters $\bar{\gamma}_{g=1}$ and $\bar{\gamma}_{g=0}$ average the CATEs over all unemployed in the respective group.

Figure 3 reports effect heterogeneity of JSP participation on cumulated months employed during the first six months after the start of participation by low and high values of the characteristics of unemployed persons. The groups in the top of Figure 3 show the largest effect heterogeneities. For groups at the bottom of Figure 3 we find only little effect heterogeneity. We estimate the largest degree of effect heterogeneity for unskilled workers. The

average effect of unskilled unemployed is 0.26 months (≈ 8 days) longer than for unemployed persons in other skill categories (see Table E.2 in Online Appendix E).

Figure 3: CATEs on cumulated employment during the first 6 months after start JSP participation by characteristics of unemployed persons.



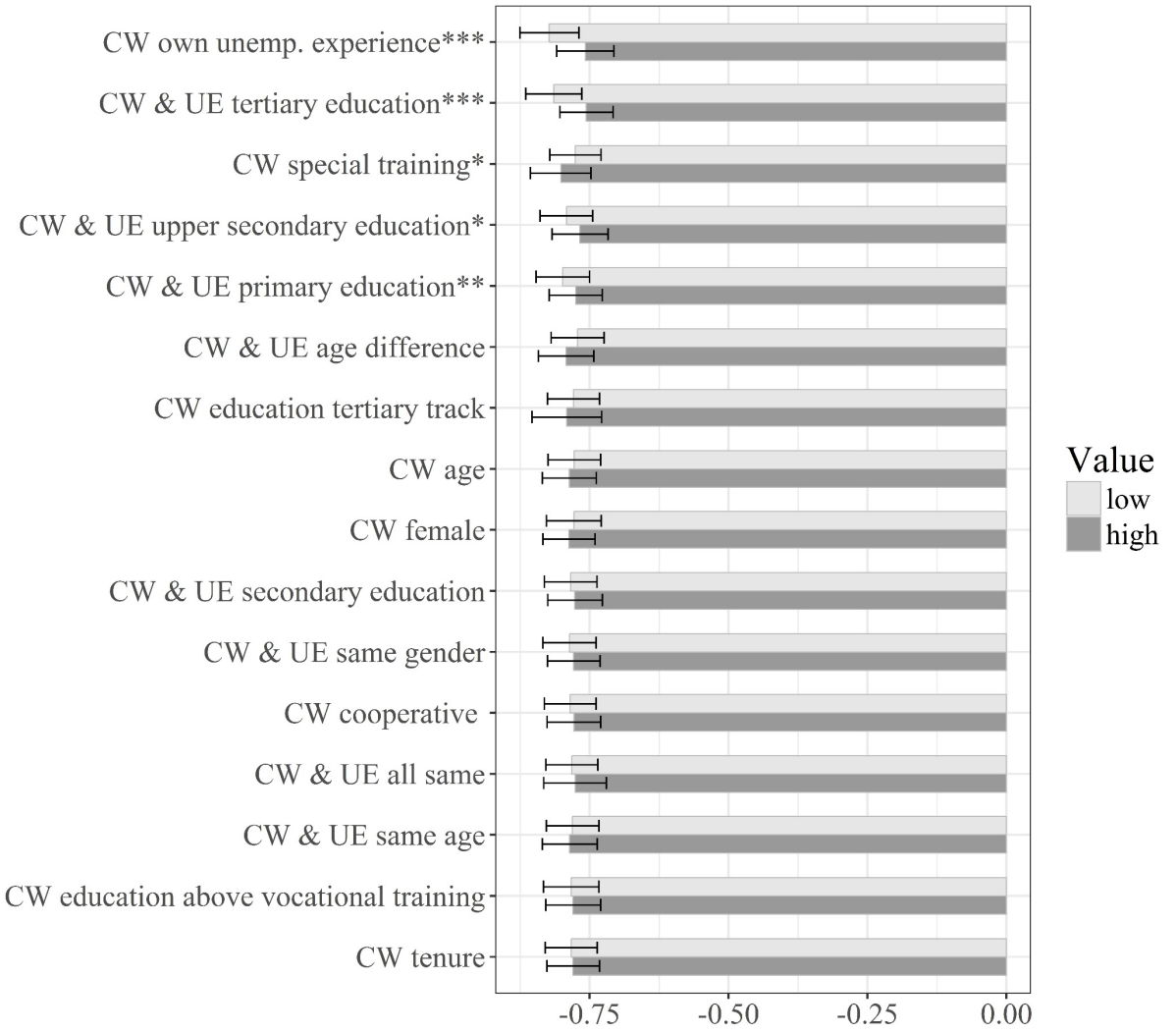
Note: CATEs by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. The CATEs are based on 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. The differences and respective standard errors are reported in Table E.2 in Online Appendix E. We report results for additional outcomes in Figures E.1, E.3, and E.5 in Online Appendix E.

Conversely, Figure 3 documents that individuals with some degree of education suffer on average more from JSP than individuals with no degree. In general, we observe that the negative lock-in effect is much less pronounced for unemployed persons with lesser qualifications. This suggests that cream-skimming reduces the effectiveness of JSP

participation. These findings are consistent with the evaluation literature (e.g., Card, Kluve, and Weber, 2015, van den Berg and van der Klaauw, 2006). Furthermore, the lock-in effects are less negative for foreigners. One potential explanation is that foreigners have a relatively small network for an informal job search. Therefore, the formal job search strategy might be relatively successful for them. This suggests more foreigners should be assigned to JSPs. We find only little heterogeneity by gender and age, which is in line with the findings of Vikström, Rosholm, and Svarer (2013) for JSPs in Denmark. In our application, the effect heterogeneity by gender is not statistically significant (see standard errors in Table E.2 in Online Appendix E).

Figure 4 reports effect heterogeneity of JSP participation on cumulated months employed during the first six months after the start of participation by low and high values of caseworker characteristics. The interpretation of Figure 4 corresponds to the interpretation of Figure 3. Although we find some significant differences, they are much less pronounced than for the characteristics of unemployed persons. Most effect heterogeneity is observed by caseworkers' own unemployment experience, but the difference is only 0.07 months ($\approx$ 2 days). However, the difference is statistically significant (see Table E.3 in Online Appendix E). Interestingly, the cooperativeness of caseworkers has no statistically significant influence on the effectiveness of JSP participation. We would have expected this characteristic to be a good predictor for effect heterogeneity, because it might approximate different monitoring intensities of the caseworker.

Figure 4: CATEs on cumulated employment during the first 6 months after start JSP participation by caseworker characteristics.



Note: CATEs by low and high values of the respective caseworker characteristic. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. The CATEs are based on 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation.. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. The differences and respective standard errors are reported in Table E.3 in Online Appendix E. We report results for additional outcomes in Figures E.2, E.4, and E.6 in Online Appendix E. CW is the abbreviation for caseworker.

5.5 Assignment rules for JSP

Next, we investigate the characteristics of unemployed persons with positive CATEs (Table 5). The number of individuals with positive CATEs amounts to 674, which corresponds to 0.9% of the unemployed persons in the sample. The first row of Table 5 reports the share of unemployed persons assigned to a JSP by the sign of the CATEs. Only 8% of unemployed

persons with positive CATEs participate in a JSP, whereas 16% of the unemployed persons with negative CATEs participate in a JSP. This points to the potential to improve the selection of JSP participants. Additionally, Table 5 reports characteristics of unemployed persons with positive and negative CATEs. The difference gives explicit advice on how assignment rules to JSPs could be improved. For example, unemployed persons with a lower past income and lesser past employment experience tend to have positive effects from participation. Participants with lower degrees of education and foreigners seem to have a higher probability to profit from a JSP. Strikingly, those unemployed persons who receive a low employability rating by their caseworker are more likely to experience positive effects from a JSP than unemployed persons with a medium or high rating. These results are further evidence that cream-skimming does not improve JSP effectiveness.

Furthermore, we document the effectiveness of hypothetical statistical assignment rules in Table 6. Statistical assignment rules have already received considerable attention in the context of ALMPs (see, e.g., Bell and Orr, 2002, Caliendo, Hujer, and Thomsen, 2008, Frölich, 2008, Dehejia, 2005, O'Leary, Decker, and Wandner, 2002, among many others). However, we are not aware of any application using machine learning methods to investigate assignment rules for ALMP that systematically consider a high-dimensional covariate space.

For the proposed assignment rules, we keep the number of 12,712 (hypothetical) JSP participants constant.[16] Therefore, the proposed assignment rules are (almost) cost neutral compared to the existing assignment mechanism. However, we do not account for possible capacity limits in regional training centres. We consider five hypothetical assignment rules: (i) random allocation (called 'random' in the following), (ii) assignment of unemployed persons with the highest CATEs (called 'best case' in the following), (iii) assignment of unemployed persons

[16] We consider only participants on the common support. Therefore, the number of participants considered here is lower than previous numbers.

with the lowest CATEs (called 'worst case' in the following), (iv) all unemployed persons with at least one unemployment spell in the previous two years and unskilled plus a random selection of the remaining unemployed persons with at least one unemployment spell in the previous two years and no degree (called 'previous unemployment' in the following), and (v) all unemployed with low employability rating by their caseworkers plus a random sample with medium employability rating (called 'employability rating' in the following). The random adding of participants in assignment rules (iv) and (v) enables us to maintain the number of 12,712 participants. The 'previous unemployment' rule (iv) is inspired by the variables that show the highest treatment effects in Table 3 and Figure 3. The 'employability rating' rule (v) assigns unemployed persons to a JSP for whom the caseworkers give a low employability rating, as opposed to cream-skimming, which assigns more unemployed persons with high employability ratings.

Table 6 reports the average CATE under the different assignment rules. The average CATE represents the hypothetical ATET under this treatment assignment. The 'worst case' and 'best case' assignment rules are the lower and upper bounds of the ATET (for a fixed number of 12,712 participants). The difference between the lower and upper bounds are about 0.65 employment months (≈ 20 days). The ATET under random assignment is -0.78 months (≈ -24 days) employment during the first six months after the start of participation. This is the benchmark assignment rule. Any imposed assignment rule should be better than random assignment. However, the observed ATET is -0.82 months (≈ -25 days) employment during the first six months after the start of the programme. It appears that the current assignment mechanism is not better than a random assignment rule. In the context of Swiss ALMPs, Lechner and Smith (2007) also find that the allocation by caseworkers performs no better than random assignment. Furthermore, this is consistent with the findings of Bell and Orr (2002) and Frölich (2008), who reject the idea that caseworkers allocate training programs efficiently in

the US and Sweden. Applying the optimal assignment rule ‘best case’ would reduce the negative employment effects by 60% ($= \big((0.82 - 0.33)/0.82\big) \cdot 100\%$).

For the proposed assignment rule ‘previous employment’ the predicted ATET is -0.51 months (≈ -15 days) employment during the first six months after the start of participation. On average, each participant has 9 days more employment under this assignment rule than under random assignment. The negative employment effect of the current assignment mechanism would be reduced by 38% ($= \big((0.82 - 0.51)/0.82\big) \cdot 100\%$). For the proposed assignment rule ‘employability rating’ the predicted ATET is -0.61 months (≈ -18 days) employment during the first six months after the start of participation. On average, each participant has 6 days more employment under this assignment rule than under random assignment. The negative employment effect of the current assignment mechanism would be reduced by 21% ($= \big((0.82 - 0.61)/0.82\big) \cdot 100\%$). These results are consistent with the argument that assignments based on expected treatment effects rather than on predicted outcomes can be more successful (Ascarza, 2016). However, the average effects remain negative and the programme does not seem useful in improving employment opportunities of unemployed persons in general. Nevertheless, the easy-to-implement assignment rules document the potential to improve the current allocation mechanism.

Table 5: Characteristics of unemployed by the sign of CATE.

	$\overline{\gamma}_i \geq 0$	$\overline{\gamma}_i < 0$	Difference	S.E.
	(1)	(2)	(3)	(4)
JSP participation	0.07	0.16	-0.09***	(0.01)
Female	0.41	0.45	-0.04	(0.07)
Past income (in 10,000 CHF)	0.32	0.42	-0.11***	(0.02)
Fraction of months emp. in last 2 years	0.70	0.80	-0.10***	(0.01)
# of unemp. spells in last 2 years	4.71	0.54	4.17***	(0.50)
Unskilled	0.62	0.24	0.38***	(0.09)
Semiskilled	0.16	0.16	0.00	(0.05)
Skilled without degree	0.14	0.04	0.10*	(0.06)
Some educational degree	0.08	0.57	-0.49***	(0.03)
Foreigner with mother tongue is cantons' language	0.14	0.11	0.03	(0.02)
Low employability rating by CW	0.43	0.14	0.29***	(0.11)
Medium employability rating by CW	0.56	0.76	-0.20*	(0.11)
High employability rating by CW	0.01	0.10	-0.10***	(0.003)
Age (in 10 years)	3.57	3.67	-0.10	(0.08)
Foreigner with temporary residence permit	0.32	0.13	0.19***	(0.06)
Foreigner with permanent residence permit	0.41	0.25	0.16**	(0.07)
# of individuals	674	77,824		

Note: Average characteristics of individuals with positive and negative CATE in the first 6 months after start of participation. The CATEs are based on 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. CW is the abbreviation for caseworker.

Table 6: Average CATE of hypothetical participants under different assignment rules.

Assignment rule	CATE for participants
	(1)
Observed (ATET)	-0.82
(i) Random	-0.78
(ii) Best case	-0.33
(iii) Worst case	-1.07
(iv) Previous unemployment	-0.51
(v) Employability rating	-0.61
# of participants	12,712

Note: Based on average predicted individual effects of 30 replications with one step efficiency augmented 10-fold cross-validated Post-Lasso.

5.6 Sensitivity checks

We perform large-scale sensitivity analyses to investigate the robustness of our results with respect to the choice of the empirical method and the selection of tuning parameters. We replicate our estimates using different forms of efficiency augmentation (see Online Appendix F.2).

As an alternative variables selector, we consider the adaptive LASSO (Zou, 2006, see Online Appendix F.3). Furthermore, we replicate the results with the Modified Outcome Method (MOM) (Signorovitch, 2007, Zhang et al., 2012, see Online Appendix F.1) instead of the MCM. Moreover, we employ radius-matching with bias adjustment (Lechner, Miquel, and Wunsch, 2011) to balance the observable covariates between the treatment and control group instead of the IPW weights. This method shows good finite sample performance (Huber, Lechner, and Wunsch, 2014). Furthermore, we compare the robustness of the main results with two different sets of additional confounders (see Online Appendix F.5 for a description how we select the additional confounders). Finally, we compare our results with the causal forest approach (Wager and Athey, 2017, see Online Appendix F.4).

Table 7 reports the correlation between the CATEs for different empirical procedures. No matter which specification we use, the correlation between the CATEs is always positive and mostly above 0.5. The causal forest CATEs are less strongly correlated, but they still show a decently strong positive association. Accordingly, our main findings are not sensitive to the choice of empirical methods or selection of tuning parameters. We report additional sensitivity checks in Online Appendix F.6. The estimation results are widely consistent across a variety of different methodological choices and estimation procedures.

Table 7: Correlation between CATEs obtained from different empirical procedures.

Cumulated employment during first 6 months	(1)	(2)	(3)	(4)	(5)	(6)
(1) MCM, one-step EA, Post-LASSO	1.00					
(2) MCM, two-step EA, Post-LASSO	0.87	1.00				
(3) MCM, no EA, Post-LASSO	0.77	0.77	1.00			
(4) MCM, one-step EA, adaptive LASSO	0.78	0.55	0.62	1.00		
(5) MCM, two-step EA, adaptive LASSO	0.77	0.56	0.58	0.87	1.00	
(6) MCM, no EA, adaptive LASSO	0.67	0.56	0.83	0.67	0.62	1.00
(7) MOM, Post-LASSO	0.75	0.77	0.81	0.58	0.56	0.64
(8) MOM, adaptive LASSO	0.62	0.49	0.66	0.67	0.72	0.71
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.97	0.87	0.76	0.74	0.73	0.66
(10) MCM, one-step EA, LASSO	0.85	0.61	0.65	0.93	0.79	0.62
(11) Procedure (1) + additional confounders 1	0.83	0.75	0.65	0.59	0.62	0.54
(12) Procedure (11) + additional confounders 2	0.90	0.86	0.72	0.62	0.65	0.59
(13) Causal forest	0.55	0.47	0.46	0.47	0.44	0.50
Cumulated employment during first 6 months	(7)	(8)	(9)	(10)	(11)	(12)
(8) MOM, adaptive LASSO	0.55	1.00				
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.74	0.61	1.00			
(10) MCM, one-step EA, LASSO	0.62	0.58	0.81	1.00		
(11) Procedure (1) + additional confounders 1	0.69	0.51	0.82	0.67	1.00	
(12) Procedure (11) + additional confounders 2	0.77	0.55	0.91	0.70	0.92	1.00
(13) Causal forest	0.43	0.40	0.55	0.51	0.50	0.52

Note: Correlations of CATEs for different methods of efficiency augmentation, variable selection, modifications and weights. EA is the abbreviation for efficiency augmentation. If not specified differently, IPW weights are used to balance the covariates. Only in procedure (9), we use radius-matching weights (Lechner Miquel, and Wunsch, 2011). See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12). Tables F.2-F.4 in Online Appendix F contain the correlation between CATEs for the other outcomes.

6 Conclusion

We investigate recently developed machine learning methods to uncover systematically treatment effect heterogeneity. We apply these methods to estimate the heterogeneous effects of Swiss Job Search Programmes (JSPs) on different employment outcomes by allowing for a high-dimensional set of variables potentially related to effect heterogeneity. We develop easy-to-implement, efficiency-improving assignment rules for JSPs.

The employment effects of JSPs are negative during the first six months after the start of participation and taper off afterwards. Parallel to this finding, we discover substantial effect heterogeneity during the first six months after the start of participation, but not afterwards.

While an appropriate assignment rule could substantially decrease the negative lock-in effects, the negative effects are unlikely to disappear completely. In particular, we find that unemployed persons with low employment opportunities as well as foreigners experience less negative effects. The data used contains the caseworkers' subjective employability rating of their clients. Using this measure alone for programme assignment, i.e. if caseworkers assign mainly unemployed persons with a low employability rating, then negative lock-in effects are already reduced by approximately 22%. The results remain consistent across a range of alternative estimators and different implementation choices, showing the robustness of the findings.

There are still many open questions that are, however, beyond the scope of this paper. On the substantive side, for example, it is not clear that the largely negative results will generalize to other economic environments and other versions of JSPs implemented in other times and other countries. On the methodological side, it must be acknowledged that despite the extensive robustness checks, these methods are still very new and there could be practical problems not yet uncovered. We investigate the heterogeneous employment effects of a particular programme for different unemployed persons. The study abstracts from the questions about an optimal programme for a particular unemployed person, which is also relevant because of the usually rich programme structure of ALMPs. Such a modified goal raises several additional statistical issues that may be addressed in future research.

References

Abbring, J.H., G.J. van den Berg (2003): "The Non-Parametric Identification of Treatment Effects in Duration Models", *Econometrica*, 71, 1491-1517.

Abbring, J.H., G.J. van den Berg (2004): "Analyzing the Effect of Dynamically Assigned Treatments using Duration Models, Binary Treatment Models, and Panel Data Models", *Empirical Economics*, 29, 5-20.

Ascarza, E. (2016): "Retention Futility: Targeting High Risk Customers might be Ineffective", *Colombia Business School Research Paper*, 16-28..

Athey, S., G.W. Imbens (2016): “Recursive Partitioning for Heterogeneous Causal Effects”, *Proceedings of the National Academy of Science of the United States of America*, 113 (27), 7353-7360.

Athey, S., G.W. Imbens (2017a): “The Econometrics of Randomized Experiments”, in *Handbook of Field Experiments*, ed. by A.V. Banerjee, E. Duflo, 1, 73-140, Elsevier, Amsterdam.

Athey, S., G.W. Imbens (2017b): “The State of Applied Econometrics - Causality and Policy Evaluation”, *Journal of Economic Perspectives*, 31 (2), 3-32.

Athey, S., S. Wager (2017): “Efficient Policy Learning”, *Working Paper*, arXiv: 1702.02896.

Behncke, S., M. Frölich, M. Lechner (2010a): “Unemployed and their Caseworkers: Should they be Friends or Foes?” *Journal of the Royal Statistical Society, Series A*, 173 (1), 67-92.

Behncke, S., M. Frölich, M. Lechner (2010b): “A Caseworker like Me – Does the Similarity between the Unemployed and their Caseworkers Increase Job Placements?” *Economic Journal* , 120 (549), 1430-1459.

Bell, S., L. Orr (2002): “Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker Home Health Aide Demonstration”, *Labour Economics*, 9 (2), 279-302.

Belloni, A., V. Chernozhukov, C. Hansen (2013): “Inference on Treatment Effects after Selection amongst High-Dimensional Controls”, *Review of Economic Studies*, 81 (2), 608-650.

Belloni, A., V. Chernozhukov, C. Hansen (2014): “High-Dimensional Methods and Inference on Structural and Treatment Effects”, *Journal of Economic Perspectives*, 28 (2), 29–50.

Berger, M., D. Black, J. Smith (2000): “Evaluating Profiling as a Means of Allocating Government Services," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, F. Pfeiffer, 59-84. Physica, Heidelberg.

Biewen, M., B. Fitzenberger, A. Osikominu, and M. Paul (2014): “The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices”, *Journal of Labor Economics*, 32 (4), 837-897.

Blasco, S, M. Rosholm (2011): “The Impact of Active Labour Market Policy on Post-Unemployment Outcomes: Evidence from a Social Experiment in Denmark”, *IZA Discussion Paper*, No. 5631.

Blundell, R., M.C. Dias, C. Meghir, J.V. Reenen (2004): “Evaluating the Employment Impact of a Mandatory Job Search Program”, *Journal of the European Economic Association*, 2 (4), 569-606.

Breiman, L. (1996): “Bagging Predictors”, *Machine Learning*, 24 (2), 123–140.

Bühlmann, P., S. van de Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, Heidelberg.

Busso, M., J. DiNardo, J. McCrary (2014): “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators”, *Review of Economics and Statistics*, 96 (5), 885–897.

Caliendo, M., R. Hujer, S. Thomsen (2008): “Identifying Effect Heterogeneity to Improve the Efficiency of Job Creation Schemes in Germany”, *Applied Economics*, 40 (9), 1101-1122.

Card, D, J. Kluve, A. Weber (2010): “Active Labour Market Policy Evaluations: A Meta Analysis”, *Economic Journal*, 120 (548), F452-F477.

Card, D, J. Kluve, A. Weber (2015): “What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations”, *Journal of the European Economic Association*, forthcoming.

Casey, K., R. Glennerster, E. Miguel (2012): “Reshaping Institutions: Evidence on Aid Impacts Using a Pre-analysis Plan”, *Quarterly Journal of Economics*, 124 (4), 1755-1812.

Chen, S., L. Tian, T. Cai, M. Yu (2017): “A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring”, *Biometrics*, forthcoming.

Chetverikov, D., Z. Liao, V. Chernozhukov (2017): “On Cross-Validated Lasso”, *Working Paper*, arXiv: 1605.02214.

Ciarleglio, A., E. Petkova, R.T. Ogden, T. Tarpey (2015): “Treatment Decisions Based on Scalar and Functional Baseline Covariates”, *Biometrics*, 71 (4), 884–894.

Cottier, L., P. Kempeneers, Y. Flückiger, R. Lalive (2017): “Does Intensive Job Search Assistance Help Job Seekers Find and Keep Jobs?”, *Working Paper*.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, P. Zamora (2013): “Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment”, *Quarterly Journal of Economics*, 128 (2), 531-80.

Crépon B, G. van den Berg (2016): “Active Labor Market Policies“, *Annual Review of Economics*, 8, 521-546.

Dehejia, R. (2005): “Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125 (1-2), 141-173.

Dolton, P., D. O’Neill (2002): “The Long-Run Effects of Unemployment Monitoring and Work-Search Programs: Experimental Evidence from the United Kingdom”, *Journal of Labor Economics*, 20 (2), 381-403.

Federal Statistical Office (2016): "Gross Domestic Product per Capita", www.bfs.admin.ch.

Fithian, W., D. Sun, J. Taylor (2017): "Optimal Inference After Model Selection", *Working Paper*, arXiv: 1410.2597.

Fredriksson, P. and Johannsen, P. (2008): "Dynamic Treatment Assignment - The Consequences for Evaluations using Observational Data", *Journal of Business and Economic and Statistics*, 26 (4), 435-445.

Frölich, M. (2008): "Statistical Treatment Choice: An Application to Active Labour Market Programmes", *Journal of the American Statistical Association*, 103 (482), 547-558.

Foster, J.C., J.M.G. Taylor, S.J. Ruberg (2011): "Subgroup Identification from Randomized Clinical Trial Data", *Statistics in Medicine*, 30 (24), 2867-2880.

Gautier P., P. Muller, B. van der Klaauw, M. Rosholm, M. Svarer (2017): "Estimating Equilibrium Effects of Job Search Assistance", *Working Paper*.

Gerfin M., M. Lechner (2002): "A Microeconometric Evaluation of Active Labour Market Policy in Switzerland", *Economic Journal*, 112 (482), 854-893.

Graversen, B.K., J.C. Van Ours (2008): "How to Help Unemployed find Jobs Quickly: Experimental Evidence from a Mandatory Activation Program" *Journal of Public Economics*, 92 (10-11), 2020-2035.

Green, D.P., H.L. Kern (2012): "Modelling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees", *Public Opinion Quarterly*, 76 (3), 491-511.

Grimmer, J., S. Messing, S.J. Westwood (2016), "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods", *Working Paper*.

Heckman, J. and Navarro, S. (2007): "Dynamic Discrete Choice and Dynamic Treatment Effects", *Journal of Econometrics*, 136 (2), 341-396.

Hirano, K., G.W. Imbens, G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*, 71 (4), 1161-1189.

Hoerl, A. E., and R. W. Kennard (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12 (1) 55-67.

Horowitz, J. L. (2015): "Variable Selection and Estimation in High-Dimensional Models", *Canadian Journal of Economics*, 48 (2), 389-407.

Horvitz, D.G., D.J. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Universe", *Journal of the American Statistical Association*, 47 (260), 663-685.

Huber, M., M. Lechner, G. Mellace (2017): “Why Do Tougher Caseworkers Increase Employment? The Role of Programme Assignment as a Causal Mechanism”, *Review of Economics and Statistics*, 99 (1), 180-183.

Huber, M., M. Lechner, C. Wunsch (2014): “The Performance of Estimators Based on the Propensity Score“, *Journal of Econometrics*, 175 (1), 1-21.

Imai, K., M. Ratkovic (2013): “Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation”, *Annals of Applied Statistic*s, 7 (1), 443-470.

Imai, K., A. Strauss (2011): “Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Applications to the Optimal Planning of the Get-Out-of-the-Vote Campaign”, *Political Analysis*, 19 (1), 1-19.

Imbens, G.W., J.M. Wooldridge (2009): “Recent Developments in the Econometrics of Program Evaluation”, *Journal of Economic Literature*, 47 (1), 5-86.

Lalive, R., J.C. van Ours, J. Zweimüller (2008): “The Impact of Active Labor Market Programs on the Duration of Unemployment”, *Economic Journal*, 118 (525), 235-257.

Lan, W., P. Zhong, R. Li, H. Wang, C. Tsai (2016): “Testing a Single Regression Coefficient in High Dimensional Linear Models”, *Journal of Econometrics*, 195 (1), 154-168.

Lechner, M. (1999): “Earnings and Employment Effects of Continuous Off-the-job Training in East Germany after Unification”, *Journal of Business Economics and Statistics*, 17 (1), 74-90.

Lechner, M. (2009): “Sequential Causal Models for the Evaluation of Labor Market Program”, *Journal of Business and Economic Statistics*, 27 (1), 71–83.

Lechner, M., R. Miquel, C. Wunsch (2011): “Long-Run Effects of Public Sponsored Training in West Germany“, *Journal of the European Economic Association*, 9 (4), 742-784.

Lechner, M., J.A. Smith (2007): “What is the Value Added by Caseworkers?” *Labour Economics*, 14 (2), 135-151.

Lechner M., A. Strittmatter (2017): “Practical Procedures to Deal with Common Support Problems in Matching Estimation”, *Econometric Reviews*, forthcoming.

Lechner, M., C. Wunsch (2009): “Are Training Programs more Effective when Unemployment is High?”, Journal of *Labor Economics*, 27 (4), 653-692.

Lechner, M., C. Wunsch (2013): “Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables”, *Labour Economics*, 21 (C), 111-121.

List, J. Shaikh, A., Y. Xu (2016): "Multiple Hypothesis Testing in Experimental Economics", *NBER Working Paper No. 21875.*

Meyer, B.D. (1995): "Lessons from the US Unemployment Insurance Experiments", *Journal of Economic Literature*, 33 (1), 91-131.

Morlok, M., D. Liechti, R. Lalive. A. Osikominu, J. Zweimüller (2014): "Evaluation der arbeitsmarktlichen Massnahmen: Wirkung auf Bewerbungsverhalten und –chancen", *SECO Publikationen, Arbeitsmarktpolitik* No. 41.

Mullainathan, S., J. Spiess (2017): "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31 (2), 87–106.

O'Leary, C., P. Decker, S. Wandner (2002): "Targeting Reemployment Bonuses", in *Targeting Employment Services*, ed. by R. Eberts, C. O'Leary, S. Wandner, 161-182, W.E. Upjohn Institute for Employment Research, Kalamazoo.

Olken, B. (2015): "Promises and Perils of Pre-Analysis Plans", *Journal of Economic Perspectives*, 29 (3), 61-80.

Qian, M., S.A. Murphy (2011): "Performance Guarantees for Individualized Treatment Rules", *Annals of Statistics*, 39 (2), 11-80.

Rinaldo, A., L. Wasserman, M. G'Sell, J. Lei, R. Tibshirani (2016). "Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference", *Working Paper*, arXiv: 1611.05401.

Robins, J.M. (1986): "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect", *Mathematical Modelling*, 7 (9-12), 1393–1512.

Rosenbaum, P.R., D.B. Rubin (1983): "The Central Role of Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70 (1), 41-55.

Rosholm, M. (2008): "Experimental Evidence on the Nature of the Danish Employment Miracle", IZA Discussion Paper No. 3620.

Rubin, D.B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66 (5), 688-701.

SECO, State Secretary for Economic Affairs (2017): "Die Lage auf dem Arbeitsmarkt im Februar 2017", www.seco.admin.ch.

Sianesi, B. (2004): "An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s", *Review of Economics and Statistics*, 86 (1), 133-155.

Signorovitch, J.E. (2007): "Identifying Informative Biological Markers in High-Dimensional Genomic Data and Clinical Trials", *PhD thesis, Harvard University*.

Su, X., C.L, Tsai, H. Wang, D.M. Nickerson, B. Li (2009): "Subgroup Analysis via Recursive Partitioning", *Journal of Machine Learning Research*, 10, 141-158.

Taddy, M., M. Gardner, L. Chen, D. Draper (2015): "A Nonparametric Bayesian Analysis of Heterogeneous Treatment Effects in Digital Experimentation", *Journal of Business and Economic Statistics*, forthcoming.

Tian, L., A.A. Alizadeh, A.J. Gentles, R. Tibshirani (2014): "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates", *Journal of the American Statistical Association*, 109 (508), 1517-1532.

Tibshirani, R. (1996): "Regression Shrinkage via the Lasso", *Journal of the Royal Statistical Society. Series B*, 58 (1), 267-288.

Van den Berg, G.J., B. van der Klaauw (2006): "Counseling and Monitoring of Unemployed Workers: Theory and Evidence from a Controlled Social Experiment", *International Economic Review*, 47 (3), 895-936.

Vansteelandt, S., T.J. VanderWeele, E.J. Tchetgen, J.M. Robins (2008): "Multiply Robust Inference for Statistical Interactions", *Journal of the American Statistical Association*, 103 (484), 1693–1704.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28 (2), 3–28.

Vikström, J., M. Rosholm, M. Svare (2013): "The Relative Efficiency of Active Labour Market Policies: Evidence from a Social Experiment and Non-Parametric Methods", *Labour Economics*, 24, 58-67.

Wager, S., S. Athey (2017): "Estimation and Inference of Heterogenous Treatment Effects using Random Forests", *Journal of the American Statistical Association*, forthcoming.

Wang, H., C. Leng (2007): "Unified LASSO Estimation by Least Squares Approximation", *Journal of the American Statistical Association*, 102 (479), 1039-1048.

Wunsch, C., M. Lechner (2008): "What Did All the Money Do? On the General Ineffectiveness of Recent West German Labour Market Programmes", *Kyklos*, 61 (1), 134-174.

Xu, Y., M. Yu, Y.-Q. Zhao, Q. Li, S. Wang, J. Shao (2015): "Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects," *Biometrics*, 71 (3), 645–653.

Zhang, B., A.A. Tsiatis, E.B. Laber, M. Davidian (2012): “A Robust Method for Estimating Optimal Treatment Regimes”, *Biometrics*, 68 (4), 1010–1018.

Zhao, Y., D. Zeng, E.B. Laber, M.R. Kosorok (2015): “New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes”, *Journal of the American Statistical Association*, 110 (510), 583-598.

Zhao, Y., D. Zeng, A.J. Rush, M.R. Kosorok (2012): “Estimating Individualized Treatment Rules using Outcome Weighted Learning” *Journal of the American Statistical Association*, 107 (449), 1106–1118.

Zou, H. (2006): “The Adaptive Lasso and its Oracle Properties”, *Journal of the American Statistical Association*, 101 (476), 1418-1429.

Online Appendices

Appendix A: Sample selection

Table A.1 documents the sample selection steps. Additional observation are trimmed to ensure common support as shown in Figure D.1.

Table A.1: Sample selection steps.

Selection criteria		Remaining sample size
Population: all new jobseekers during the year 2003		238,902
Exclude Geneva and five other employment offices	-19,464	219,438
Exclude jobseekers not (yet) assigned to a caseworker	-4,289	215,149
Exclude foreigners with work permit shorter than one year	-5,399	209,750
Exclude jobseekers without unemployment benefit claim	-18,434	191,316
Exclude jobseekers who applied for or claim disability insurance	-3,163	188,153
Restrict to prime-age population (24 to 55 years old)	-51,649	136,504
Exclude unemployed whose caseworker did not respond to the questionnaire	-31,469	105,035
Exclude unemployed whose caseworkers did not respond to the cooperativeness question	-4,915	100,120
Exclude participants in other ALMP than JSP	-8,787	91,333
Exclude individuals employed at (pseudo) treatment date	-6,135	85,198

Note: Only the last two sample selection steps differ from Huber, Lechner, Mellace (2017).

Appendix B: Descriptive statistics

The following table shows unconditional means and standard deviations by participation status as well as standardised differences to illustrate selection into participation.

Table B.1: Descriptive statistics of confounding variables by JSP participation status.

	Participants		Non-Participants		Std. Diff.
	Mean	S.D.	Mean	S.D.	
	(1)	(2)	(3)	(4)	(5)
Characteristics of unemployed persons					
Female	0.45	-	0.44	-	0.56
× French speaking REA	0.04	-	0.11	-	19.51
× Italian speaking REA	0.01	-	0.04	-	11.85
Age (in 10 years)	3.73	0.88	3.66	0.86	5.60
Unskilled	0.22	-	0.23	-	1.80
× French speaking REA	0.03	-	0.05	-	8.36
× Italian speaking REA	0.01	-	0.03	-	8.62
Semi-skilled qualification	0.15	-	0.16	-	2.45
× French speaking REA	0.02	-	0.05	-	12.10
× Italian speaking REA	0.002	-	0.01	-	5.16
Skilled qualification without degree	0.03	-	0.05	-	4.72
× French speaking REA	0.003	-	0.02	-	11.22
× Italian speaking REA	0.002	-	0.01	-	4.11
Employability rating low	0.12	-	0.14	-	3.98
× French speaking REA	0.01	-	0.02	-	9.87
× Italian speaking REA	0.004	-	0.01	-	4.94
Employability rating medium	0.77	-	0.74	-	5.80
× French speaking REA	0.07	-	0.19	-	26.32
× Italian speaking REA	0.02	-	0.05	-	11.57
# of unemp. spells in last 2 years	0.41	0.98	0.64	1.27	13.86
× French speaking REA	0.05	0.36	0.19	0.76	16.84
× Italian speaking REA	0.02	0.22	0.07	0.46	10.16
# of emp. spells in last 5 years	0.10	0.13	0.13	0.15	14.70
Fraction of months emp. in last 2 years	0.83	0.22	0.79	0.25	12.57
× French speaking REA	0.06	0.22	0.19	0.35	30.04
× Italian speaking REA	0.02	0.13	0.06	0.22	15.77
Past income (in 10,000 CHF)	4.58	2.02	4.16	2.05	14.50
Prev. job in primary sector	0.06	-	0.10	-	10.44
Prev. job in secondary sector	0.16	-	0.13	-	6.04
Prev. job in tertiary sector	0.63	-	0.58	-	7.07
Prev. job self-employed	0.004	-	0.01	-	3.01
Prev. job manager	0.08	-	0.07	-	1.85
Prev. job skilled worker	0.63	-	0.60	-	4.70
Prev. job unskilled worker	0.26	-	0.29	-	5.01

Table continues on next page >

Table B.1 continued.

	Participants		Non-Participants		Std. Diff.
	Mean	S.D.	Mean	S.D.	
	(1)	(2)	(3)	(4)	(5)
Characteristics of unemployed persons					
Native language not German, French, or Italian	0.29	-	0.32	-	5.40
× French speaking REA	0.02	-	0.08	-	18.01
× Italian speaking REA	0.01	-	0.02	-	9.80
Married	0.47	-	0.49	-	2.35
Foreigner with temporary residence permit	0.11	-	0.14	-	6.96
Foreigner with permanent residence permit	0.23	-	0.25	-	3.12
Foreigner with mother tongue similar to canton's language	0.12	-	0.11	-	2.40
Lives in big city	0.17	-	0.17	-	0.05
Lives in medium sized city	0.16	-	0.13	-	4.83
Start of JSP participation in the second unemp. quarter	0.45	-	0.46	-	0.38
Caseworker characteristics					
Female	0.45	-	0.41	-	6.94
× French speaking REA	0.02	-	0.09	-	22.33
× Italian speaking REA	0.01	-	0.02	-	6.15
Age (in 10 years)	4.43	1.16	4.44	1.16	0.77
× French speaking REA	0.37	1.29	1.14	2.04	31.79
× Italian speaking REA	0.11	0.70	0.34	1.19	16.43
Tenure (in years)	5.54	3.23	5.86	3.31	6.84
× French speaking REA	0.47	1.78	1.59	3.07	31.36
× Italian speaking REA	0.21	1.39	0.60	2.29	14.58
Own unemp. experience	0.63	-	0.63	-	0.54
× French speaking REA	0.05	-	0.17	-	26.33
× Italian speaking REA	0.02	-	0.05	-	11.73
Education above vocational training	0.45	-	0.43	-	2.36
× French speaking REA	0.04	-	0.10	-	17.68
× Italian speaking REA	0.01	-	0.03	-	9.46
Education tertiary track	0.21	-	0.24	-	4.68
× French speaking REA	0.02	-	0.09	-	21.92
× Italian speaking REA	0.004	-	0.02	-	8.25
Vocational training degree	0.26	-	0.23	-	5.63
× French speaking REA	0.002	-	0.01	-	9.64
× Italian speaking REA	0.01	-	0.04	-	11.28
Indicator for missing caseworker characteristics	0.04	-	0.04	-	0.13

Table continues on next page >

Table B.1 continued.

	Participants		Non-Participants		Std. Diff.
	Mean	S.D.	Mean	S.D.	
	(1)	(2)	(3)	(4)	(5)
Allocation of unemployed persons to caseworkers					
By industry	0.66	-	0.53	-	17.73
× French speaking REA	0.05	-	0.10	-	12.88
× Italian speaking REA	0.01	-	0.04	-	11.32
By occupation	0.58	-	0.56	-	3.08
× French speaking REA	0.06	-	0.17	-	25.14
× Italian speaking REA	0.01	-	0.05	-	14.27
By age	0.04	-	0.03	-	2.58
By employability	0.07	-	0.07	-	0.12
By region	0.09	-	0.12	-	7.55
Other allocation type	0.07	-	0.07	-	1.37
Local labour market characteristics					
French speaking REA	0.08	-	0.25	-	33.30
Italian speaking REA	0.03	-	0.08	-	16.81
Cantonal unemployment rate (in %)	3.64	0.77	3.75	0.86	9.23
× French speaking REA	0.32	1.10	1.05	1.86	33.93
× Italian speaking REA	0.11	0.69	0.34	1.16	16.61
Cantonal GDP per capita (in 10,000 CHF)	5.13	0.92	4.92	0.93	15.75
# of caseworker	989		1,282		
# of observations	12,998		72,200		

Note: We report unconditional means for all variables, standard deviations (S.D.) for all non-binary variables, and standardised differences between participants and non-participants. Rosenbaum and Rubin (1983) consider a standardised difference of more than 20 as being 'large'. We report the descriptive statistics of the outcome variables in Table 1 of the main text. REA is the abbreviation for regional employment agency. For many variables we include interactions with a dummy for French (× French speaking REA) and Italian (× Italian speaking REA) speaking regional employment agencies. To account for categorical variables, we omit the dummies some qualification degree, employability rating high, lives in small city, and German speaking regional employment agencies.

Appendix C: Proof of Theorem 1

The following proof is based on the seminal contributions of Rosenbaum and Rubin (1983). To proof the identification of CATEs, we use the definition (see Section 4.1 in the main text),

$$\gamma(z) = E\left[Y_i^1 - Y_i^0 \middle| Z_i = z\right] = E\left[Y_i^1 \middle| Z_i = z\right] - \left[Y_i^0 \middle| Z_i = z\right].$$

Then we apply the law of iterative expectations

$$\gamma(z) = E_{X|Z=z}\left[E\left[Y_i^1 \middle| X_i = x, Z_i = z\right] \middle| Z_i = z\right] - E_{X|Z=z}\left[E\left[Y_i^0 \middle| X_i = x, Z_i = z\right] \middle| Z_i = z\right].$$

When we condition on the confounders X_i, the potential outcomes are independent of the treatment indicator D_i (Assumption 1),

$$\begin{aligned}\gamma(z) = {} & E_{X|Z=z}\left[E\left[Y_i^1 \middle| D_i = 1, X_i = x, Z_i = z\right] \middle| Z_i = z\right] \\ & - E_{X|Z=z}\left[E\left[Y_i^0 \middle| D_i = 0, X_i = x, Z_i = z\right] \middle| Z_i = z\right].\end{aligned}$$

Conditional on the treatment status, the potential outcomes equal the observed outcome,

$$\begin{aligned}\gamma(z) = {} & E_{X|Z=z}[E[Y_i | D_i = 1, X_i = x, Z_i = z] | Z_i = z] \\ & - E_{X|Z=z}[E[Y_i | D_i = 0, X_i = x, Z_i = z] | Z_i = z].\end{aligned}$$

Appendix D: Propensity score and matching quality

Table D.1 reports the average marginal effects of the propensity score estimation to illustrate selection into participation. Most of the significant coefficients confirm the observation that unemployed with higher skills and labour market success are more likely to participate in the program.

Table D.2 shows then that inverse probability weighting successfully balances the covariates indicated by a maximum standardised difference of 2.44 and a mean standardised difference of 0.7.

Table D.1: Average marginal effects in the propensity score estimation.

	Av. Marg. Eff.	S.E.
	(1)	
Characteristics of unemployed persons		
Female	0.01	(0.004)
× French speaking REA	0.01	(0.01)
× Italian speaking REA	-0.03	(0.02)
Age (in 10 years)	-0.01	(0.01)
Age^2/10,000	0.21	(0.17)
Unskilled	0.01*	(0.01)
× French speaking REA	0.10***	(0.02)
× Italian speaking REA	0.05**	(0.02)
Semi-skilled qualification	0.002	(0.01)
× French speaking REA	0.06***	(0.01)
× Italian speaking REA	0.03	(0.03)
Skilled qualification without degree	0.01*	(0.01)
× French speaking REA	-0.03	(0.03)
× Italian speaking REA	0.02	(0.03)
Employability rating low	-0.04***	(0.01)
× French speaking REA	0.10***	(0.03)
× Italian speaking REA	0.13***	(0.03)
Employability rating medium	-0.02	(0.01)
× French speaking REA	0.09***	(0.02)
× Italian speaking REA	0.07***	(0.02)
# of unemp. spells in last 2 years	-0.01***	(0.002)
× French speaking REA	0.003	(0.004)
× Italian speaking REA	0.004	(0.01)
Number of emp. spells in last 5 years	-0.08***	(0.01)
Fraction of months emp. in last 2 years	0.03***	(0.01)
× French speaking REA	-0.03	(0.02)
× Italian speaking REA	-0.05*	(0.02)
Past income (in 10,000 CHF)	0.09***	(0.01)
Prev. job in primary sector	-0.04***	(0.01)
Prev. job in secondary sector	0.04***	(0.01)
Prev. job in tertiary sector	0.01**	(0.01)
Prev. job self-employed	-0.09***	(0.02)
Prev. job manager	-0.05***	(0.01)
Prev. job skilled worker	-0.02**	(0.01)
Prev. job unskilled worker	-0.02**	(0.01)
Native language not German, French, or Italian	-0.01**	(0.01)
× French speaking REA	-0.03**	(0.01)
× Italian speaking REA	-0.01	(0.02)
Married	0.002	(0.003)
Foreigner with temporary residence permit	-0.02***	(0.01)
Foreigner with permanent residence permit	0.002	(0.004)
Foreigner with mother tongue similar to canton's language	0.03***	(0.004)
Lives in big city	-0.01	(0.01)
Lives in medium sized city	0.02***	(0.01)
Start JSP participation in second unemp. quarter	0.02***	(0.004)

Table continues on next page >

Table D.1 continued.

	Av. Marg. Eff. (1)	S.E.
Caseworker characteristics		
Female	0.02**	(0.01)
× French speaking REA	-0.05*	(0.03)
× Italian speaking REA	0.05*	(0.02)
Age (in 10 years)	0.003	(0.004)
× French speaking REA	0.001	(0.001)
× Italian speaking REA	0.001	(0.001)
Tenure (in years)	0.002	(0.001)
× French speaking REA	-0.01	(0.01)
× Italian speaking REA	0.003	(0.004)
Own unemp. experience	0.01	(0.01)
× French speaking REA	-0.03	(0.03)
× Italian speaking REA	0.05	(0.03)
Education above vocational training	0.0001	(0.01)
× French speaking REA	0.01	(0.03)
× Italian speaking REA	0.01	(0.03)
Education tertiary track	0.002	(0.01)
× French speaking REA	-0.02	(0.04)
× Italian speaking REA	0.01	(0.04)
Vocational training degree	0.02*	(0.01)
× French speaking REA	-0.09	(0.07)
× Italian speaking REA	0.02	(0.03)
Indicator for missing caseworker characteristics	0.002	(0.02)
Allocation of unemployed to caseworkers		
By industry	0.02***	(0.01)
× French speaking REA	0.05*	(0.03)
× Italian speaking REA	-0.04*	(0.02)
By occupation	0.02**	(0.01)
× French speaking REA	0.04*	(0.03)
× Italian speaking REA	-0.06**	(0.03)
By age	0.01	(0.01)
By employability	-0.02	(0.01)
By region	-0.04**	(0.01)
Other allocation type	-0.03**	(0.01)
Local labour market characteristics		
French speaking REA	-0.06	(0.09)
Italian speaking REA	-0.19	(0.11)
Cantonal unemployment rate (in %)	0.03***	(0.01)
× French speaking REA	-0.07***	(0.01)
× Italian speaking REA	-0.03*	(0.02)
Cantonal GDP per capita (in 10,000 CHF)	-0.03***	(0.01)
# of caseworker	1,282	
# of observations	85,198	

Note: The estimation is based on a Probit model with the outcome JSP participation. The Probit model includes a constant term. We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. REA is the abbreviation for regional employment agency. For many variables we include interactions with a dummy for French (× French speaking REA) and Italian (× Italian speaking REA) speaking regional employment agencies. To account for categorical variables, we omit the dummies some qualification degree, employability rating high, lives in small city, and German speaking regional employment agencies.

Table D.2: Balance of confounders after matching.

	Participants	Non-Participants	Std. Diff.
	Mean	Mean	
	(1)	(2)	(3)
Characteristics of unemployed persons			
Female	0.44	0.44	0.21
× French speaking REA	0.08	0.08	0.01
× Italian speaking REA	0.03	0.03	0.66
Age (in 10 years)	3.65	3.67	1.57
$Age^2/10{,}000$	0.14	0.14	1.59
Unskilled	0.24	0.24	0.31
× French speaking REA	0.05	0.05	0.56
× Italian speaking REA	0.02	0.02	0.44
Semi-skilled qualification	0.16	0.16	0.27
× French speaking REA	0.03	0.04	1.43
× Italian speaking REA	0.01	0.01	0.45
Skilled qualification without degree	0.04	0.04	1.42
× French speaking REA	0.01	0.01	1.30
× Italian speaking REA	0.01	0.01	1.46
Employability rating low	0.15	0.14	1.00
× French speaking REA	0.01	0.01	0.06
× Italian speaking REA	0.01	0.01	0.03
Employability rating medium	0.75	0.75	0.02
× French speaking REA	0.14	0.15	0.70
× Italian speaking REA	0.04	0.04	0.20
# of unemp. spells in last 2 years	0.59	0.57	1.01
× French speaking REA	0.12	0.11	0.36
× Italian speaking REA	0.05	0.05	0.68
# of emp. spells in last 5 years	0.12	0.12	1.19
Fraction of months emp. in last 2 years	0.80	0.80	0.48
× French speaking REA	0.13	0.13	0.83
× Italian speaking REA	0.05	0.05	1.23
Past income (in 10,000 CHF)	4.21	4.24	0.85
Prev. job in primary sector	0.08	0.08	0.45
Prev. job in secondary sector	0.14	0.14	0.17
Prev. job in tertiary sector	0.59	0.60	0.72
Prev. job self-employed	0.01	0.01	0.40
Prev. job manager	0.07	0.07	0.10
Prev. job skilled worker	0.59	0.60	1.19
Prev. job unskilled worker	0.30	0.30	0.89

Table continues on next page >

Table D.2 continued.

	Participants	Non-Participants	Std. Diff.
	Mean	Mean	
	(1)	(2)	(3)
Characteristics of unemployed persons			
Native language not German, French, or Italian	0.32	0.32	0.23
× French speaking REA	0.05	0.05	0.50
× Italian speaking REA	0.02	0.02	0.30
Married	0.48	0.48	0.36
Foreigner with temporary residence permit	0.13	0.13	0.06
Foreigner with permanent residence permit	0.25	0.25	0.61
Foreigner with mother tongue similar to canton's language	0.11	0.11	0.10
Lives in big city	0.16	0.17	0.98
Lives in medium sized city	0.15	0.14	1.09
Start of JSP participation in the second unemp. quarter	0.43	0.45	1.72
Caseworker characteristics			
Female	0.41	0.41	0.35
× French speaking REA	0.05	0.05	0.15
× Italian speaking REA	0.02	0.02	0.15
Age (in 10 years)	4.43	4.43	0.24
× French speaking REA	0.79	0.80	0.50
× Italian speaking REA	0.27	0.29	0.92
Tenure (in years)	5.77	5.75	0.29
× French speaking REA	1.09	1.09	0.22
× Italian speaking REA	0.48	0.52	1.26
Own unemp. experience	0.63	0.63	0.36
× French speaking REA	0.11	0.11	0.46
× Italian speaking REA	0.04	0.04	0.71
Education above vocational training	0.44	0.45	0.95
× French speaking REA	0.08	0.08	0.50
× Italian speaking REA	0.02	0.02	1.73
Education tertiary track	0.23	0.22	1.41
× French speaking REA	0.06	0.06	2.44
× Italian speaking REA	0.01	0.01	0.18
Vocational training degree	0.23	0.24	0.40
× French speaking REA	0.01	0.01	0.47
× Italian speaking REA	0.03	0.04	1.36
Indicator for missing caseworker characteristics	0.04	0.04	0.24

Table continues on next page >

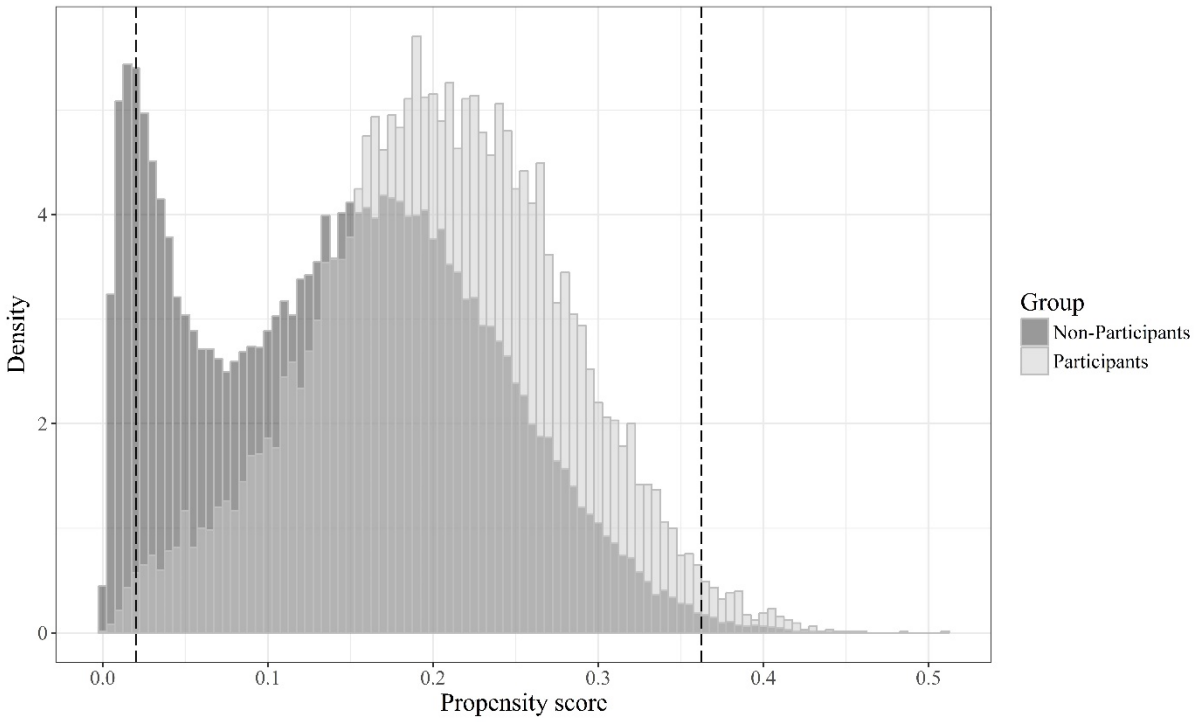
Table D.2 continued.

	Participants	Non-Participants	Std. Diff.
	Mean	Mean	
		(2)	(3)
Allocation of unemployed to caseworkers			
By industry	0.58	0.58	0.56
× French speaking REA	0.09	0.09	0.54
× Italian speaking REA	0.03	0.03	1.16
By occupation	0.56	0.56	0.01
× French speaking REA	0.13	0.13	0.13
× Italian speaking REA	0.04	0.04	1.19
By age	0.03	0.03	0.53
By employability	0.08	0.07	1.23
By region	0.11	0.11	1.31
Other allocation type	0.08	0.07	1.89
Local labour market characteristics			
French speaking REA	0.17	0.18	0.41
Italian speaking REA	0.06	0.07	1.16
Cantonal unemployment rate (in %)	3.68	3.68	0.29
× French speaking REA	0.69	0.71	1.02
× Italian speaking REA	0.27	0.29	1.10
Cantonal GDP per capita (in 10,000 CHF)	4.98	4.98	0.18
# of trimmed observations	582	6,118	
# of observations after trimming	12,712	65,786	

Note: We report IPW re-weighted means for all variables and standardised differences between participants and non-participants. Rosenbaum and Rubin (1983) consider a standardised difference of more than 20 as being 'large'. REA is the abbreviation for regional employment agency. For many variables we include interactions with a dummy for French (× French speaking REA) and Italian (× Italian speaking REA) speaking regional employment agencies. To account for categorical variables, we omit the dummies some qualification degree, employability rating high, lives in small city, and German speaking regional employment agencies.

Figure D.1 plots the distribution of the propensity score by participation status. We enforce common support by trimming observations below the 0.5 quantile the participants and above the 99.5 quantile of non-participants. In total, we trim 6,700 observations (582 participants, 6,118 non-participants). This procedure shows good final sample performance in the study Lechner and Strittmatter (2017).

Figure D.1: Histogram of the propensity score by participation status.



Note: The histogram has a binwidth of 0.005. The dashed lines show the lower and upper threshold of trimming.

Appendix E: Results for additional outcome variables

This Appendix complements the results shown in section 5.3 and 5.4 of the main text by providing results for additional outcome variables.

Table E.1: Post-LASSO coefficients of selected variables for the employment outcome 31 months after start participation.

	Months employed during first 31 months after the start of participation	
	Coef.	S.E.
	(1)	
Constant	-1.37***	(0.32)
Female × CW education above vocational training	0.07	(0.54)
Unskilled × CW education above vocational training	-1.24	(1.56)
Unskilled × prev. job unskilled	4.66***	(1.58)
# of unemp. spells in last 2 years × unemp. person and CW have primary education	0.29	(0.20)
Fraction of months emp. in last 2 years × past income 57 - 75k	0.32	(0.63)
GDP per capita × prev. job self-employed	4.56	(5.60)
CW education: above vocational training × past income 25 - 50k	1.09*	(0.66)
CW education: tertiary track × past income 25 - 50k	0.21	(0.84)
Degree in vocational training for caseworkers × past income 50 - 75k	-0.70	(0.81)
Married × past income 50 - 75k	-0.37	(0.70)
Foreigner with permanent residence permit × past income 50 - 75k	0.32	(0.76)
Medium city × prev. job unskilled	-0.48	(0.94)
Single household × no emp. spell last 2 years	-0.24	(0.56)
Past income 0 - 25k × # emp. spells past 5 years	3.60	(2.47)
# emp. spells past 5 years × unemp. person and CW have primary education	-1.41	(1.87)
# emp. spells past 5 years × unemp. person and CW have same gender, age, and education	-6.34	(5.93)
No emp. spell last 2 years × skilled worker	-0.56	(0.53)
Prev. job in primary sector × unskilled	-0.20	(1.30)
Unskilled × unemp. person and CW have primary education	0.09	(0.56)
Regional emp. agency No. 44	-0.98	(1.27)
# of selected variables	20 of 1,268	

Note: We apply one-step efficiency augmentation. We partition the data randomly into selection and estimation sample. We choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We obtain standard errors (S.E.) from a clustered bootstrap at caseworker level with 4,999 replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. CW is the abbreviation for caseworker. 25 - 50k is the abbreviation for 25,000-50,000 CHF. 50 – 75k is the abbreviation for 50,000-75,000 CHF. We omit the results for the outcome cumulated employment between months 25-31 after start of JSP participation, because we do not identify any heterogeneity variables for this outcome in the random sample spilt we consider.

Table E.2: Differences between CATEs on cumulated employment during the first 6 months after start JSP participation by characteristics of unemployed persons.

	Difference	S.E.
	(1)	(2)
Unskilled	0.26***	(0.03)
Some degree	-0.25***	(0.02)
# of unemployment spells last 2 years	0.24***	(0.03)
Employability rating low	0.23***	(0.03)
Skilled qualification w/o degree	0.21***	(0.05)
Past income	-0.19***	(0.02)
Foreigner with temporary permit	0.12***	(0.03)
Foreigner with permanent permit	0.12***	(0.02)
Employability rating high	-0.11***	(0.01)
Employability rating medium	-0.10***	(0.02)
Fraction employed last year	-0.09***	(0.01)
Semi-skilled qualification	0.05**	(0.02)
Mother tongue of canton	0.04***	(0.01)
Age	0.03***	(0.01)
Female	0.02	(0.02)

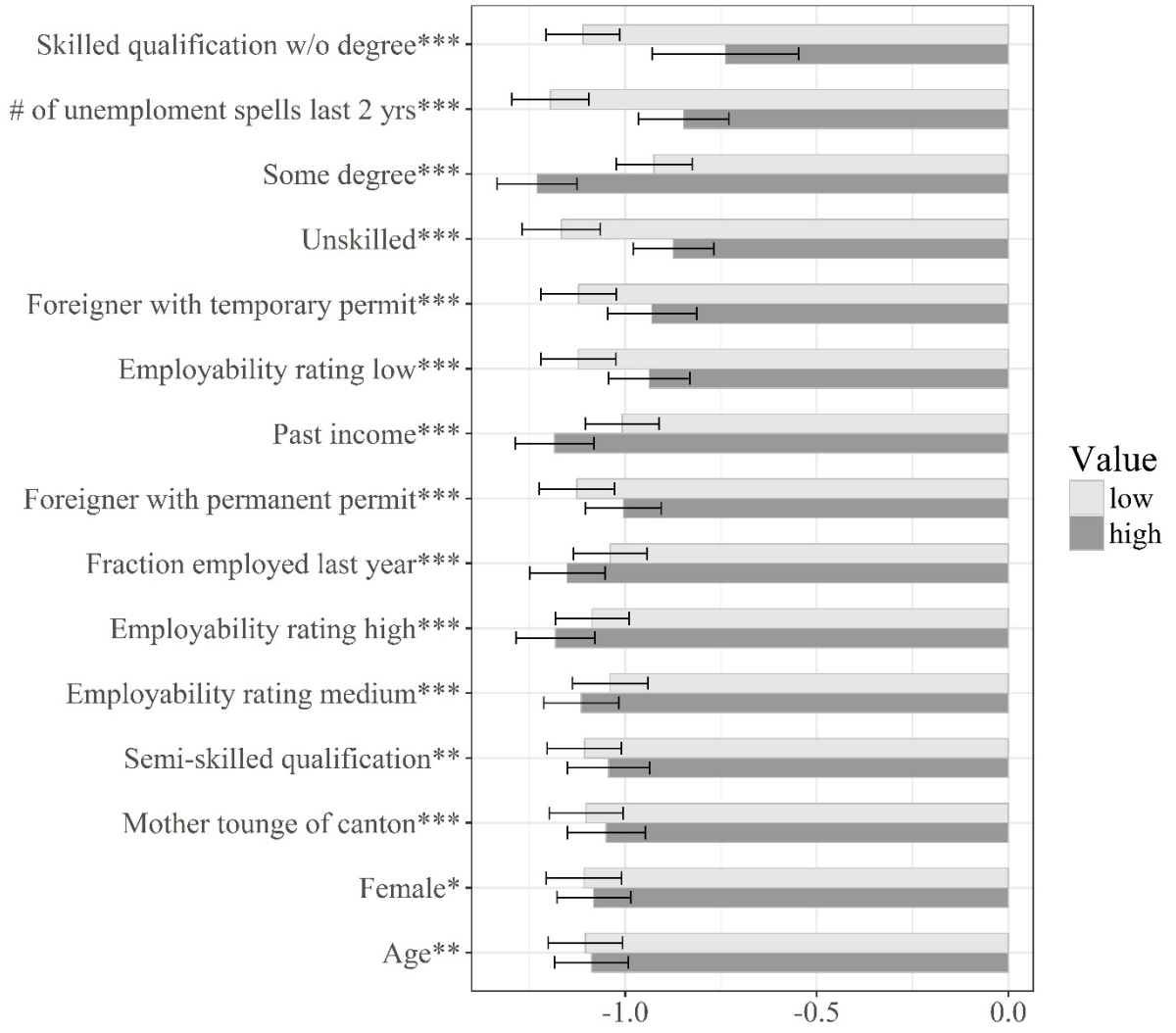
Note: This table reports the differences between CATE by low and high values of the respective characteristic of unemployed persons (see also Figure 3). A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. The CATEs are based on 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report standard errors based on 1,000 bootstrap replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Table E.3: Differences between CATEs on cumulated employment during the first 6 months after start JSP participation by caseworker characteristics.

	Difference	S.E.
	(1)	(2)
CW own unemployment experience	0.07***	(0.02)
CW & UE tertiary education	0.06***	(0.02)
CW special training	-0.03*	(0.01)
CW & UE upper secondary education	0.02*	(0.01)
CW & UE primary education	0.02**	(0.01)
CW & UE age difference	-0.02	(0.01)
CW education tertiary track	-0.01	(0.02)
CW age	-0.01	(0.01)
Female	-0.01	(0.01)
CW & UE secondary education	0.01	(0.01)
CW & UE same gender	0.01	(0.01)
CW cooperative	0.01	(0.01)
CW & UE same gender, age, and education	0.01	(0.02)
CW & UE same age ± 5 years	-0.01	(0.01)
CW education above vocational training	0.00	(0.02)
CW tenure	0.00	(0.01)

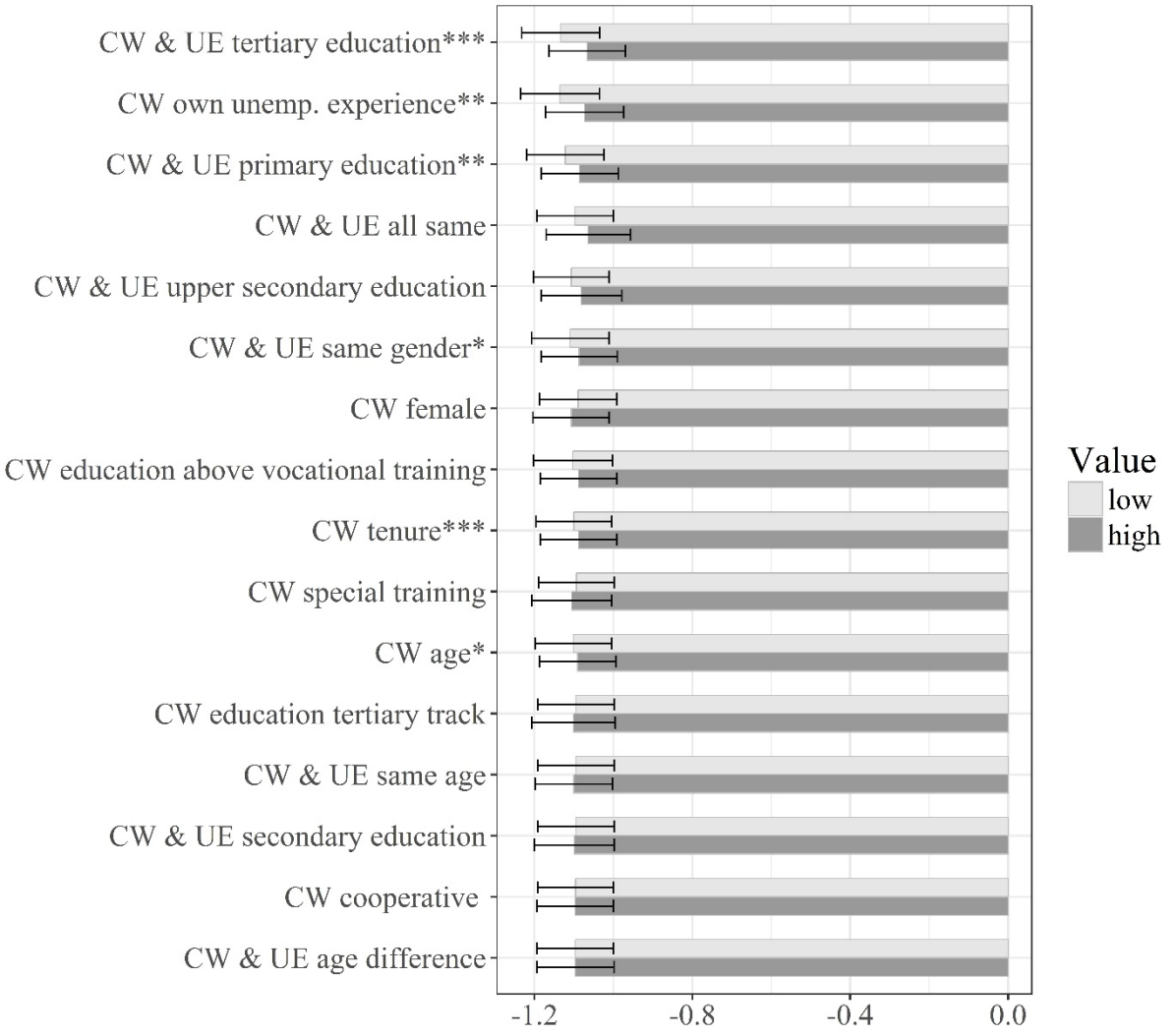
Note: This table reports the differences between CATE by low and high values of the respective characteristic of unemployed persons (see also Figure 4). A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. The CATEs are based on 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report standard errors based on 1,000 bootstrap replications. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Figure E.1: CATE on cumulated employment during the first 12 months after start JSP participation by characteristics of unemployed persons.



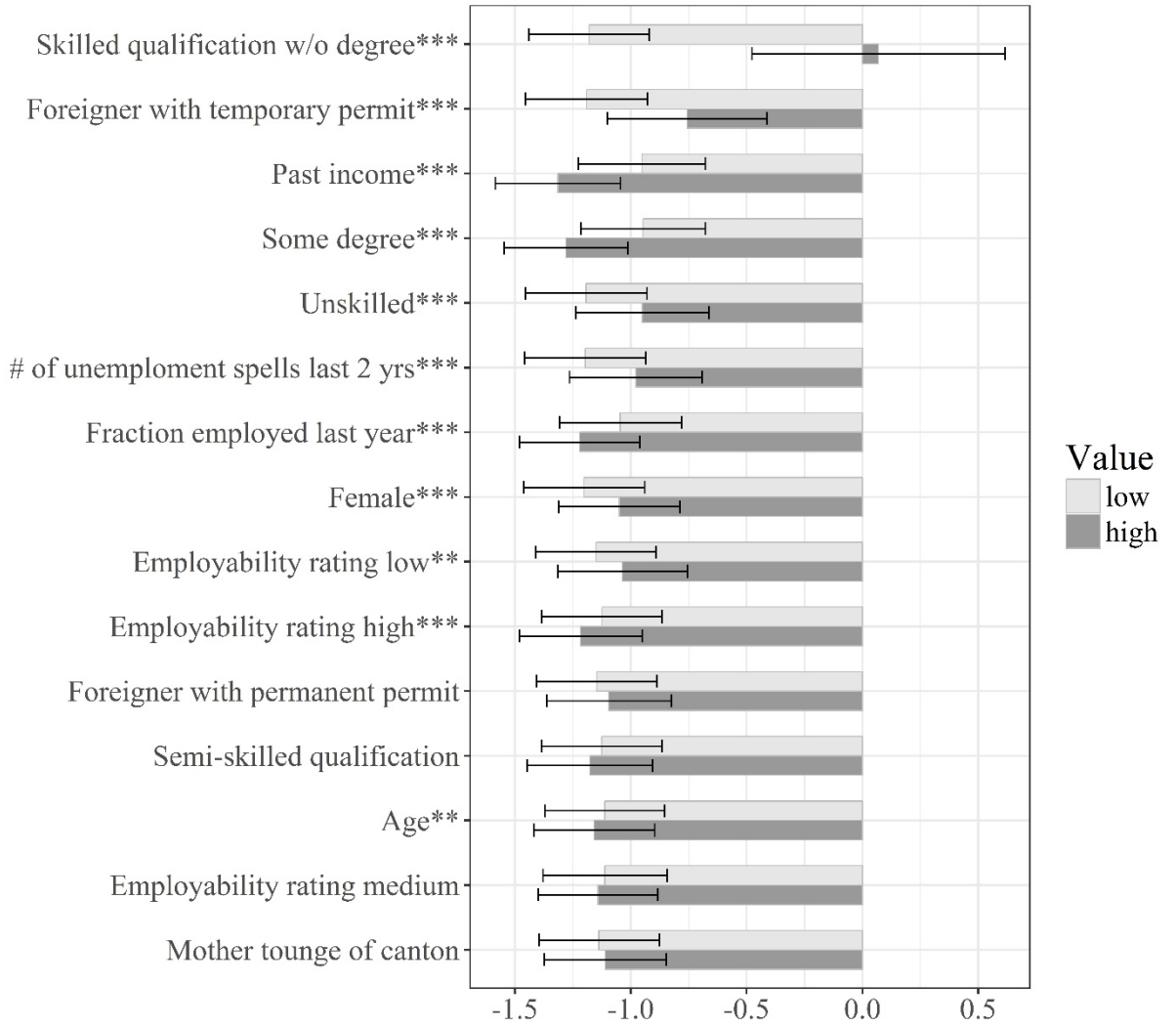
Note: CATE by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Figure E.2: CATE on cumulated employment during the first 12 months after start JSP participation by caseworker characteristics.



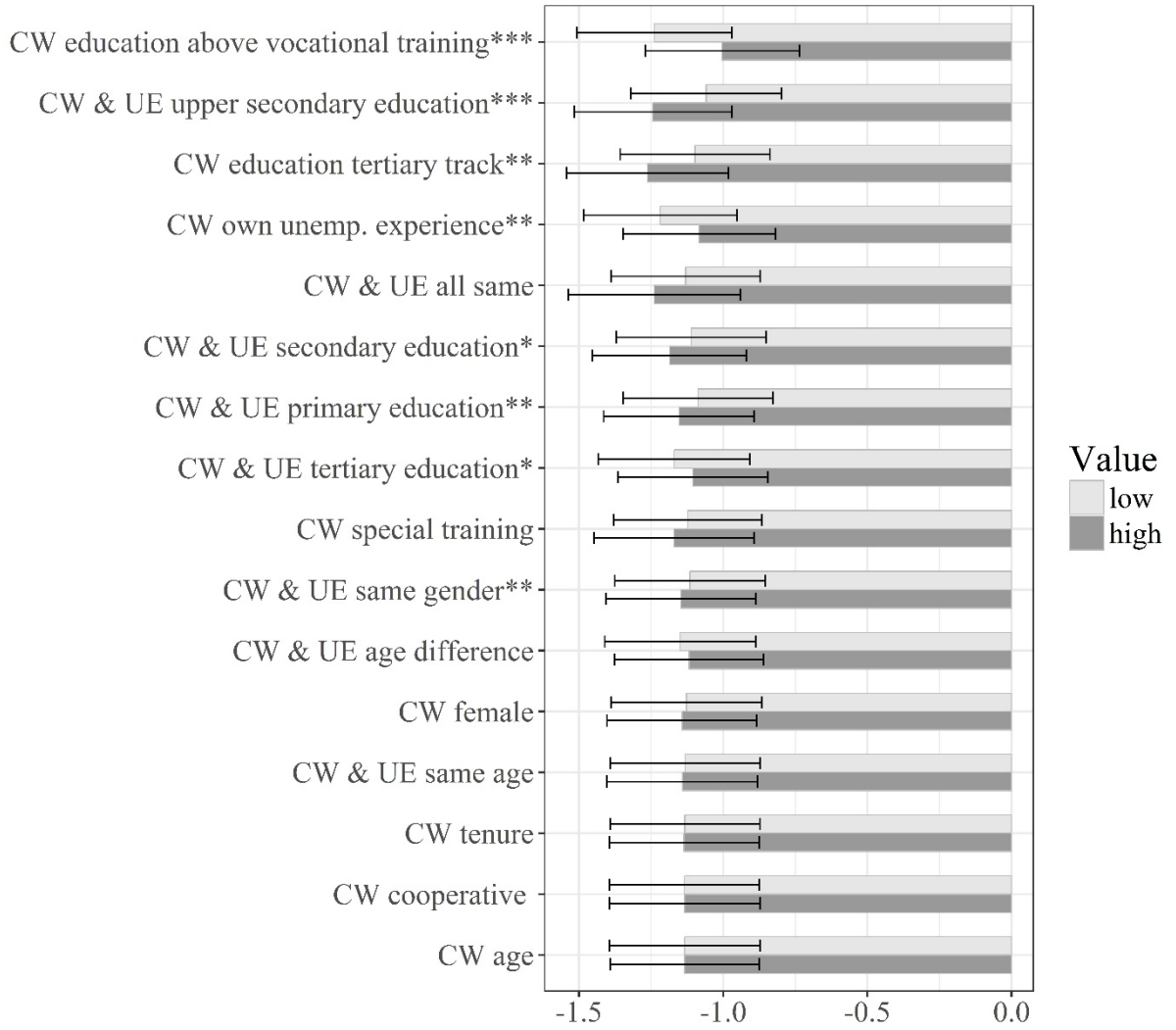
Note: CATE by low and high values of the respective caseworker characteristic. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. CW is the abbreviation for caseworker.

Figure E.3: CATE on cumulated employment during the first 31 months after start JSP participation by characteristics of unemployed persons.



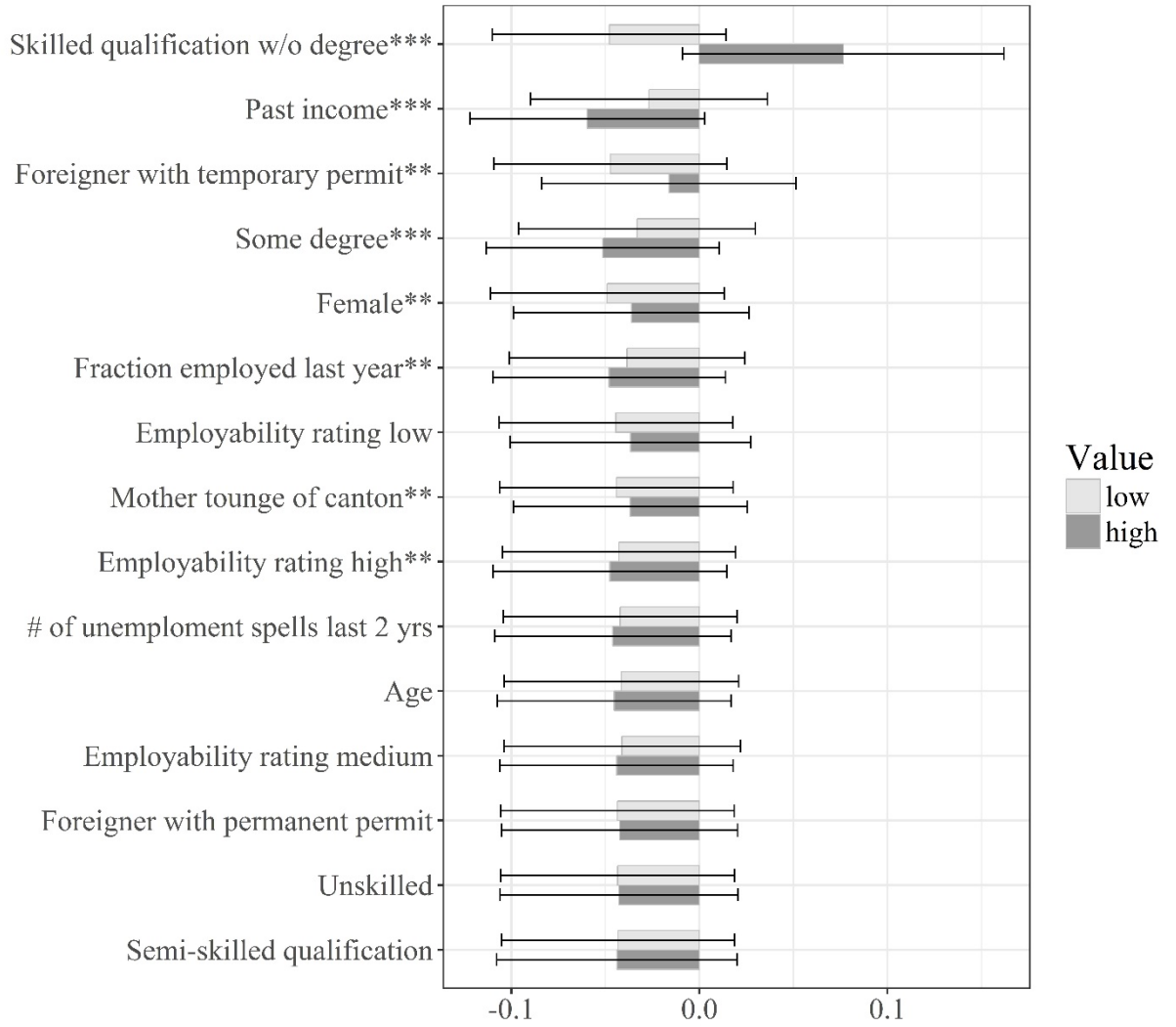
Note: CATE by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Figure E.4: CATE on cumulated employment during the first 31 months after start JSP participation by caseworker characteristics.



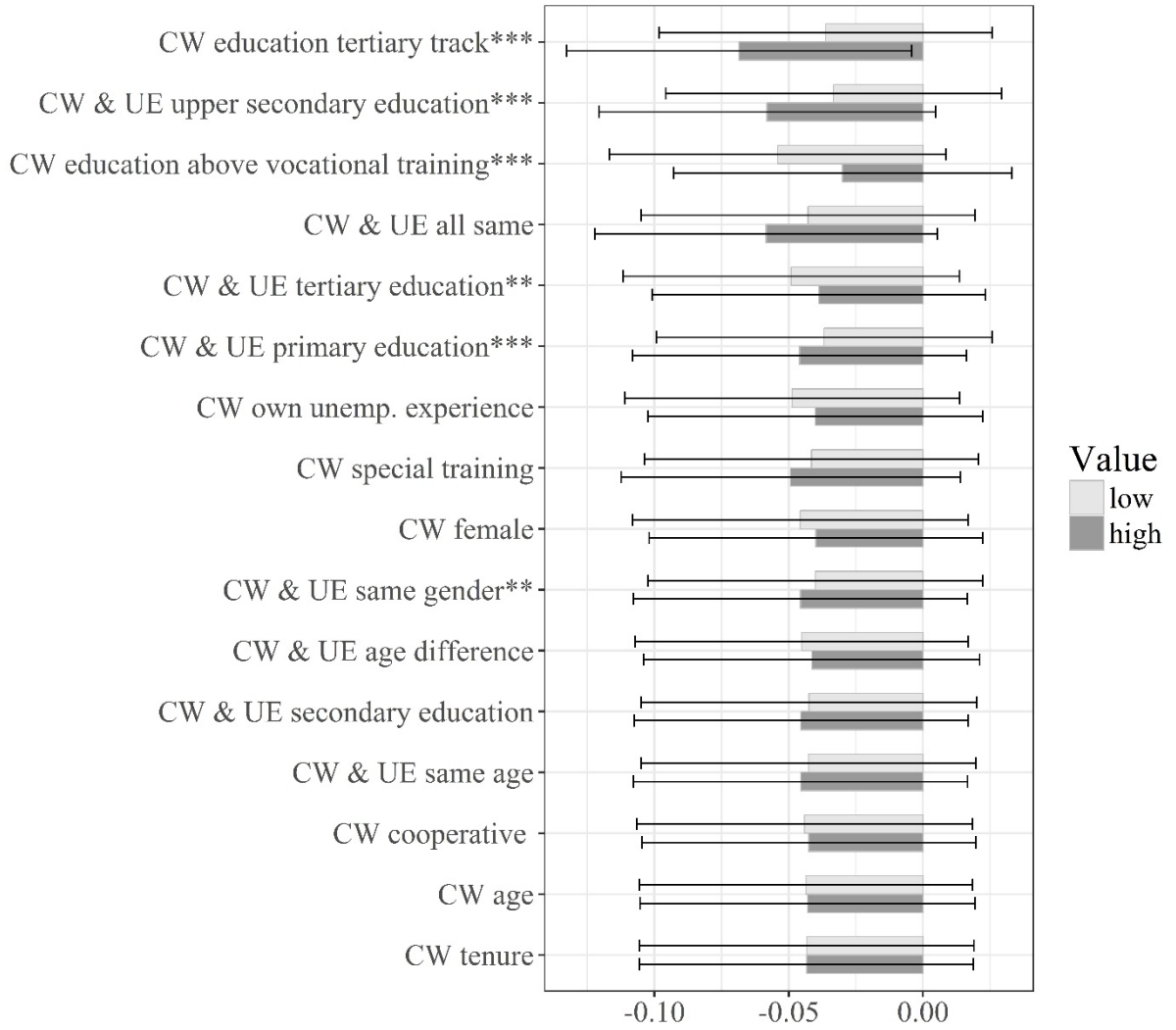
Note: CATE by low and high values of the respective caseworker characteristic. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation.. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. CW is the abbreviation for caseworker.

Figure E.5: CATE on cumulated employment during the months 25-31 after start JSP participation by characteristics of unemployed persons.



Note: CATE by low and high values of the respective characteristic of unemployed persons. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively.

Figure E.6: CATE on cumulated employment during the months 25-31 after start JSP participation by caseworker characteristics.



Note: CATE by low and high values of the respective caseworker characteristic. A low value is zero when the variable is binary or below the median when the variable is non-binary. A high value is one when the variable is binary or not below the median when the variable is non-binary. We aggregate the CATEs over 30 random sample splits. For each partition, we choose the penalty term based on Post-LASSO RMSE, which we optimise with 10-fold cross-validation. We apply one-step efficiency augmentation. We report the 95%-confidence interval based on the bootstrap procedure described in section 4.6. *, **, *** mean statistically different from zero at the 10%, 5%, 1% level, respectively. CW is the abbreviation for caseworker.

Appendix F: Robustness checks

F.1 Modified Outcome Method (MOM)

In the baseline estimations, we rely on the MCM, because it offers the possibility of efficiency augmentation described below. An alternative approach is the Modified Outcome Method (MOM), which modifies the outcome instead of the covariates. This procedure was proposed by Signorovich (2007) and extended to non-experimental studies by Zhang et al. (2012). We apply the MOM by minimising the objective function

$$\arg\min_{\hat{\delta}} \left[\sum_{i=1}^{N} \left(Y_i^* - Z_i \hat{\delta} \right)^2 + \lambda \sum_{j=1}^{p} \left| \hat{\delta}_j \right| \right],$$

where $Y_i^* = \hat{w}(D_i, X_i, Z_i) \cdot Y_i$ is the modified outcome.

F.2 Efficiency augmentation

Chen et al. (2017) propose two ways to account for the main effects, which might improve the efficiency of the selection procedure. First, the one-step procedure includes the main effects in the empirical model by solving

$$\arg\min_{\hat{\delta}, \hat{\beta}_t} \left[\sum_{i=1}^{N} \hat{w}(D_i, X_i, Z_i) T_i \left(Y_i - Z_i \hat{\beta}_t - \frac{Z_i T_i}{2} \hat{\delta} \right)^2 + \lambda \sum_{j=1}^{p} \left(\left| \hat{\beta}_{tj} \right| + \left| \hat{\delta}_j \right| \right) \right].$$

This specification is strongly related to the approach of Imai and Ratkovic (2013), but they consider only experimental research designs and propose a to use a combination of LASSO und Support Vector Machines.

Second, the two-step procedure estimates a WOLS including only the main effects in the first place. Afterwards, the estimated residuals $\hat{u}$ of this auxiliary regression are used as regressand when selecting the interaction effects

$$\arg\min_{\hat{\delta}} \left[\sum_{i=1}^{N} \hat{w}(D_i, X_i, Z_i) T_i \left(\hat{u}_i - \frac{Z_i T_i}{2} \hat{\delta} \right)^2 + \lambda \sum_{j=1}^{p} \left| \hat{\delta}_j \right| \right].$$

We consider the one-step procedure in the main specifications and show sensitivity checks with the two-step procedure.

F.3 Adaptive LASSO

In the main results, we rely on the standard LASSO estimator. A potential disadvantage of this estimator is the inability to penalize the coefficients differentially. The adaptive LASSO is an alternative estimator that received a lot of attention in the literature (see Zou, 2006). One way of specifying the adaptive LASSO in high-dimensional settings, is to minimise the objective function

$$\arg\min_{\hat{\delta}} \left[\sum_{i=1}^{N} \hat{w}(D_i, X_i, Z_i) T_i \left(Y_i - \frac{Z_i T_i}{2} \hat{\delta} \right)^2 + \sum_{j=1}^{p} \frac{\lambda}{\left| \hat{\beta}_j \right|} \left| \hat{\delta}_j \right| \right], \tag{1}$$

where we obtain $\hat{\beta}_j$ from a first step Ridge estimator minimising

$$\arg\min_{\hat{\beta}} \left[\sum_{i=1}^{N} \hat{w}(D_i, X_i, Z_i) T_i \left(Y_i - \frac{Z_i T_i}{2} \hat{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \right].$$

The Ridge estimator penalises the sum of squared coefficients instead of the sum of the absolute coefficients (Hoerl and Kennard, 1970). Therefore, Ridge estimators shrink the coefficients to zero, but they do not reach zero, unless the penalty parameter is infinity. Accordingly, Ridge estimators do not select models. The penalty term of the adaptive LASSO in equation (1) decreases with the absolute size of the Ridge coefficients. Accordingly, variables with small Ridge coefficients have a larger penalty term in the adaptive LASSO. Zou (2006) shows that the adaptive LASSO can achieve under appropriate assumptions the oracle property. The oracle property implies that the adaptive LASSO selects the correct model at an asymptotically

appropriate rate, such that the selection step can be neglected. Wang and Leng (2007) discuss the properties of the adaptive LASSO.

F.4 Causal forest

We implement the approach suggested by Wager and Athey (2017). It is based on combining the causal tree approach by Athey and Imbens (2016) with the idea of random forests. In other words, deep trees are build and effects are estimated within the resulting leaves. Tree building is based on maximising estimated heterogeneity and using a randomly selected subset of features at each possible sample split. Then, these individual predictions of the CATEs are averaged over many bootstrap samples. So far, experience with these causal random forests are limited which is the reason why we use them only for the robustness analysis.

F.5 Additional confounders

In our main specifications, we use the propensity score specification of Huber, Lechner, and Mellace (2017). We consider two additional sets of confounding variables to check the sensitivity of our results with respect to misspecification of the propensity score. First, we estimate a LASSO model on the treatment equation

$$\underset{\widehat{\alpha},\widehat{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(D_i - X_i\hat{\alpha} - Z_i\hat{\beta} \right)^2 \right] + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|,$$

where the confounders X_i are not penalised. We consider all variables of Z_i with non-zero LASSO coefficients $\hat{\beta}$ as additional confounders (we call them 'additional confounders 1' in the following). We denote the additional confounders 1 by Z_i^1.

Second, we estimate a LASSO model on the outcome equation

$$\underset{\widehat{\alpha},\widehat{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(Y_i - X_i\hat{\alpha} - Z_i^1\hat{\beta} - Z_i\hat{\gamma} \right)^2 \right] + \lambda \sum_{j=1}^{p} |\hat{\gamma}_j|,$$

where the confounders X_i and Z_i^1 are not penalised. We consider all variables of Z_i with non-zero LASSO coefficients $\hat{\gamma}$ as additional confounders (we call them 'additional confounders 2' in the following). This procedure to select additional confounders is in the spirit of double selection (see Belloni, Chernozhukov, and Hansen, 2013).

F.6 Results of additional robustness checks

We estimate CATEs, $\hat{\gamma}_s(Z_i)$, using different random sample splits. Table F.1 reports the average correlations between CATEs obtained from the different random sample splits. The CATEs are positively correlated between the different random sample splits. The positive correlations are particularly high when we consider the employment outcome during the first six month after start participation. After longer time periods, the positive correlations decrease, but remain decently positive. This suggests that our results are not sensitive to a particular random sample split. This finding is robust across 12 different estimation procedures we consider.

Tables F.2 to F.4 documents the correlations of aggregated CATEs, $\bar{\gamma}(Z_i)$, obtained across different estimation procedures. These tables are similar to Table 7, but consider different employment outcomes. The CATEs obtained from the alternative methods are highly positively correlated, no matter which procedure we use. The smallest correlations are found for the outcome cumulated employment during months 25 to 31 after the start of participation. For this outcome the correlations are substantially lower. This is not surprising, because only little heterogeneity is found for this outcome in general.

Table F.5 provides the descriptive statistics of all CATEs by different estimation procedures and outcome variables. The means are close to the respective ATEs, which is expected under the law of iterative expectations. This reassures that all estimation procedures are able to replicate the semi-parametric IPW estimates.

Interestingly, the differences in the standard deviations indicate that some estimation procedures detect more heterogeneity than others. We observe three striking patterns: First, the two-step efficiency augmentation detects less heterogeneity than the one-step efficiency augmentation or no efficiency augmentation. Second, the adaptive LASSO without efficiency augmentation finds most effect heterogeneity. Third, the procedure with the weights obtained from radius-matching tends to detect slightly less heterogeneity than the estimation procedures using IPW weights. Table F.6 describes the number of selected variables in the 30 random sample splits we consider. We observe large differences over different estimation procedure. The adaptive LASSO selects substantially more variables than Post-LASSO estimators. This could be an explanation why the adaptive LASSO detects most effect heterogeneity.

Table F.1: Average correlation between CATEs obtained from different random sample splits.

	Months employed since start participation			
	During first 6 months	During first 12 months	During first 31 months	During months 25-31
	(1)	(2)	(3)	(4)
(1) MCM, one-step EA, Post-LASSO	0.67	0.56	0.46	0.34
(2) MCM, two-step EA, Post-LASSO	0.66	0.57	0.24	0.22
(3) MCM, no EA, Post-LASSO	0.68	0.66	0.53	0.52
(4) MCM, one-step EA, adaptive LASSO	0.55	0.50	0.47	0.47
(5) MCM, two-step EA, adaptive LASSO	0.47	0.35	0.32	0.29
(6) MCM, no EA, adaptive LASSO	0.54	0.53	0.57	0.55
(7) MOM, Post-LASSO	0.63	0.64	0.46	0.44
(8) MOM, adaptive LASSO	0.50	0.47	0.44	0.41
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.64	0.52	0.30	0.24
(10) MCM, one-step EA, LASSO	0.59	0.53	0.48	0.51
(11) Procedure (1) + additional confounders 1	0.66	0.32	0.68	0.44
(12) Procedure (11) + additional confounders 2	0.66	0.47	0.54	0.40

Note: We estimate CATEs using different random sample splits and report the average correlation. We consider different methods of efficiency augmentation, variable selection, modifications and weights. EA is the abbreviation for efficiency augmentation. If not otherwise specified, IPW weights are used to balance the covariates. Only in procedure (9) we do use radius-matching weights (Lechner Miquel, and Wunsch, 2011). See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).

Table F.2: Correlation between CATEs obtained from different empirical procedures.

Cumulated employment during first 12 months	(1)	(2)	(3)	(4)	(5)	(6)
(1) MCM, one-step EA, Post-LASSO	1.00					
(2) MCM, two-step EA, Post-LASSO	0.88	1.00				
(3) MCM, no EA, Post-LASSO	0.73	0.62	1.00			
(4) MCM, one-step EA, adaptive LASSO	0.71	0.49	0.57	1.00		
(5) MCM, two-step EA, adaptive LASSO	0.75	0.52	0.54	0.89	1.00	
(6) MCM, no EA, adaptive LASSO	0.66	0.48	0.84	0.70	0.66	1.00
(7) MOM, Post-LASSO	0.61	0.52	0.68	0.44	0.40	0.54
(8) MOM, adaptive LASSO	0.62	0.46	0.72	0.69	0.71	0.76
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.93	0.84	0.73	0.64	0.64	0.66
(10) MCM, one-step EA, LASSO	0.82	0.61	0.62	0.93	0.85	0.69
(11) Procedure (1) + additional confounders 1	0.74	0.74	0.52	0.48	0.49	0.45
(12) Procedure (11) + additional confounders 2	0.83	0.91	0.57	0.50	0.53	0.47
(13) Causal forest	0.45	0.37	0.36	0.38	0.36	0.43

Cumulated employment during first 12 months	(7)	(8)	(9)	(10)	(11)	(12)
(8) MOM, adaptive LASSO	0.52	1.00				
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.59	0.61	1.00			
(10) MCM, one-step EA, LASSO	0.49	0.63	0.73	1.00		
(11) Procedure (1) + additional confounders 1	0.59	0.44	0.69	0.57	1.00	
(12) Procedure (11) + additional confounders 2	0.61	0.45	0.82	0.59	0.82	1.00
(13) Causal forest	0.29	0.40	0.45	0.43	0.35	0.36

Note: Correlations of CATEs for different methods of efficiency augmentation, variable selection, modifications and weights. EA is the abbreviation for efficiency augmentation. If not otherwise specified, IPW weights are used to balance the covariates. Only in procedure (9) we do use radius-matching weights (Lechner Miquel, and Wunsch, 2011). (11), (12) and (13) are estimated on different common support. Thus, the correlations are calculated for those observations being on support in both specifications. See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).

Table F.3: Correlation between CATEs obtained from different empirical procedures.

Cumulated employment during first 31 months	(1)	(2)	(3)	(4)	(5)	(6)
(1) MCM, one-step EA, Post-LASSO	1.00					
(2) MCM, two-step EA, Post-LASSO	0.88	1.00				
(3) MCM, no EA, Post-LASSO	0.54	0.41	1.00			
(4) MCM, one-step EA, adaptive LASSO	0.68	0.62	0.52	1.00		
(5) MCM, two-step EA, adaptive LASSO	0.82	0.80	0.46	0.79	1.00	
(6) MCM, no EA, adaptive LASSO	0.57	0.46	0.83	0.63	0.58	1.00
(7) MOM, Post-LASSO	0.46	0.33	0.65	0.37	0.34	0.50
(8) MOM, adaptive LASSO	0.57	0.47	0.69	0.70	0.61	0.72
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.88	0.67	0.57	0.59	0.61	0.59
(10) MCM, one-step EA, LASSO	0.82	0.71	0.58	0.90	0.80	0.64
(11) Procedure (1) + additional confounders 1	0.16	0.23	0.06	0.16	0.21	0.06
(12) Procedure (11) + additional confounders 2	0.24	0.26	0.13	0.21	0.23	0.12
(13) Causal forest	0.33	0.23	0.37	0.32	0.26	0.41
Cumulated employment during first 31 months	(7)	(8)	(9)	(10)	(11)	(12)
(8) MOM, adaptive LASSO	0.49	1.00				
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.49	0.53	1.00			
(10) MCM, one-step EA, LASSO	0.45	0.65	0.74	1.00		
(11) Procedure (1) + additional confounders 1	0.06	0.19	0.06	0.18	1.00	
(12) Procedure (11) + additional confounders 2	0.11	0.26	0.19	0.24	0.93	1.00
(13) Causal forest	0.26	0.41	0.39	0.36	0.04	0.12

Note: Correlations of CATEs for different methods of efficiency augmentation, variable selection, modifications and weights. EA is the abbreviation for efficiency augmentation. If not otherwise specified, IPW weights are used to balance the covariates. Only in procedure (9) we do use radius-matching weights (Lechner Miquel, and Wunsch, 2011). (11), (12) and (13) are estimated on different common support. Thus, the correlations are calculated for those observations being on support in both specifications. See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).

Table F.4: Correlation between CATEs obtained from different empirical procedures.

Cumulated employment during months 25-31	(1)	(2)	(3)	(4)	(5)	(6)
(1) MCM, one-step EA, Post-LASSO	1.00					
(2) MCM, two-step EA, Post-LASSO	0.83	1.00				
(3) MCM, no EA, Post-LASSO	0.46	0.44	1.00			
(4) MCM, one-step EA, adaptive LASSO	0.69	0.75	0.47	1.00		
(5) MCM, two-step EA, adaptive LASSO	0.72	0.81	0.43	0.84	1.00	
(6) MCM, no EA, adaptive LASSO	0.48	0.51	0.82	0.61	0.56	1.00
(7) MOM, Post-LASSO	0.47	0.44	0.89	0.45	0.38	0.72
(8) MOM, adaptive LASSO	0.55	0.59	0.67	0.73	0.66	0.74
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.81	0.68	0.47	0.62	0.54	0.46
(10) MCM, one-step EA, LASSO	0.80	0.79	0.49	0.89	0.80	0.58
(11) Procedure (1) + additional confounders 1	0.20	0.27	0.09	0.24	0.28	0.10
(12) Procedure (11) + additional confounders 2	0.20	0.26	0.11	0.23	0.32	0.12
(13) Causal forest	0.26	0.29	0.37	0.33	0.32	0.41
Cumulated employment during months 25-31	(7)	(8)	(9)	(10)	(11)	(12)
(8) MOM, adaptive LASSO	0.65	1.00				
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	0.51	0.51	1.00			
(10) MCM, one-step EA, LASSO	0.48	0.65	0.76	1.00		
(11) Procedure (1) + additional confounders 1	0.11	0.25	0.23	0.29	1.00	
(12) Procedure (11) + additional confounders 2	0.09	0.28	0.24	0.27	0.87	1.00
(13) Causal forest	0.38	0.41	0.31	0.35	0.04	0.08

Note: Correlations of CATEs for different methods of efficiency augmentation, variable selection, modifications and weights. EA is the abbreviation for efficiency augmentation. If not otherwise specified, IPW weights are used to balance the covariates. Only in procedure (9) we do use radius-matching weights (Lechner Miquel, and Wunsch, 2011). (11), (12) and (13) are estimated on different common support. Thus, the correlations are calculated for those observations being on support in both specifications. See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).

Table F.5: Descriptive statistics of CATEs by estimation procedure.

	Mean	Median	S.D.	Min	Max
	(1)	(2)	(3)	(4)	(5)
Cumulated employment during first 6 months					
(1) MCM, one-step EA, Post-LASSO	-0.78	-0.84	0.25	-1.41	0.77
(2) MCM, two-step EA, Post-LASSO	-0.76	-0.83	0.15	-0.90	0.28
(3) MCM, no EA, Post-LASSO	-0.80	-0.85	0.29	-1.95	0.93
(4) MCM, one-step EA, adaptive LASSO	-0.77	-0.81	0.32	-2.44	1.14
(5) MCM, two-step EA, adaptive LASSO	-0.77	-0.81	0.16	-1.46	0.67
(6) MCM, no EA, adaptive LASSO	-0.79	-0.81	0.26	-2.01	0.46
(7) MOM, Post-LASSO	-0.76	-0.83	0.27	-1.46	1.27
(8) MOM, adaptive LASSO	-0.77	-0.80	0.17	-1.39	1.47
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	-0.77	-0.83	0.23	-1.39	0.62
(10) MCM, one-step EA, LASSO	-0.77	-0.81	0.36	-2.33	1.26
(11) Procedure (1) + additional confounders 1	-0.80	-0.85	0.20	-1.30	1.13
(12) Procedure (11) + additional confounders 2	-0.77	-0.82	0.19	-1.22	0.71
(13) Causal forest	-0.82	-0.83	0.11	-1.36	0.15
Cumulated employment during first 12 months					
(1) MCM, one-step EA, Post-LASSO	-1.10	-1.20	0.32	-2.09	1.44
(2) MCM, two-step EA, Post-LASSO	-1.06	-1.13	0.14	-1.22	0.20
(3) MCM, no EA, Post-LASSO	-1.09	-1.14	0.54	-4.56	1.90
(4) MCM, one-step EA, adaptive LASSO	-1.06	-1.11	0.56	-4.70	3.12
(5) MCM, two-step EA, adaptive LASSO	-1.05	-1.10	0.27	-2.59	1.24
(6) MCM, no EA, adaptive LASSO	-1.08	-1.08	0.60	-4.46	1.74
(7) MOM, Post-LASSO	-0.98	-1.05	0.61	-3.59	4.64
(8) MOM, adaptive LASSO	-1.02	-1.07	0.40	-3.33	2.48
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	-1.09	-1.18	0.26	-1.93	0.87
(10) MCM, one-step EA, LASSO	-1.07	-1.14	0.60	-4.02	2.90
(11) Procedure (1) + additional confounders 1	-1.29	-1.34	0.14	-1.59	1.20
(12) Procedure (11) + additional confounders 2	-1.04	-1.10	0.18	-1.31	0.53
(13) Causal forest	-1.06	-1.06	0.22	-2.09	0.84

Table continues on next page >

Table F.5 continued.

	Mean	Median	S.D.	Min	Max
	(1)	(2)	(3)	(4)	(5)
Cumulated employment during first 31 months					
(1) MCM, one-step EA, Post-LASSO	-1.13	-1.25	0.60	-3.79	4.12
(2) MCM, two-step EA, Post-LASSO	-1.19	-1.23	0.20	-1.91	1.07
(3) MCM, no EA, Post-LASSO	-1.18	-1.19	1.39	-10.97	6.75
(4) MCM, one-step EA, adaptive LASSO	-1.06	-1.10	1.49	-11.57	11.33
(5) MCM, two-step EA, adaptive LASSO	-1.06	-1.08	0.52	-3.79	3.49
(6) MCM, no EA, adaptive LASSO	-1.19	-1.14	1.79	-9.60	10.03
(7) MOM, Post-LASSO	-0.88	-0.95	1.45	-7.02	16.17
(8) MOM, adaptive LASSO	-1.02	-1.11	1.35	-7.63	8.28
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	-1.14	-1.25	0.42	-2.82	1.62
(10) MCM, one-step EA, LASSO	-1.05	-1.11	1.26	-8.00	7.13
(11) Procedure (1) + additional confounders 1	-1.80	-1.83	0.38	-4.24	6.67
(12) Procedure (11) + additional confounders 2	-1.20	-1.26	0.32	-1.68	7.12
(13) Causal forest	-0.82	-0.83	0.59	-4.62	3.89
Cumulated employment during months 25-31					
(1) MCM, one-step EA, Post-LASSO	-0.04	-0.05	0.06	-0.32	0.48
(2) MCM, two-step EA, Post-LASSO	-0.05	-0.05	0.03	-0.27	0.26
(3) MCM, no EA, Post-LASSO	-0.03	-0.05	0.27	-1.45	1.67
(4) MCM, one-step EA, adaptive LASSO	0.00	0.00	0.29	-1.95	2.01
(5) MCM, two-step EA, adaptive LASSO	0.00	0.01	0.12	-0.77	1.53
(6) MCM, no EA, adaptive LASSO	-0.03	-0.03	0.38	-1.76	2.36
(7) MOM, Post-LASSO	0.01	-0.02	0.20	-0.77	1.53
(8) MOM, adaptive LASSO	0.00	-0.01	0.31	-1.64	2.57
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	-0.04	-0.05	0.03	-0.17	0.19
(10) MCM, one-step EA, LASSO	-0.03	-0.04	0.18	-1.26	1.27
(11) Procedure (1) + additional confounders 1	-0.03	-0.03	0.06	-1.18	0.63
(12) Procedure (11) + additional confounders 2	-0.08	-0.08	0.03	-0.16	0.44
(13) Causal forest	0.08	0.08	0.14	-0.84	0.93

Note: We obtain CATEs are based on 30 different random sample splits. Standard deviations are abbreviated with S.D. in column (3). See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).

Table F.6: Number of selected variables in different estimation procedures.

	Mean	Median	S.D.	Min	Max
	(1)	(2)	(3)	(4)	(5)
Cumulated employment during first 6 months					
(1) MCM, one-step EA, Post-LASSO	34.9	32	18.2	5	87
(2) MCM, two-step EA, Post-LASSO	5.7	4	5.6	1	24
(3) MCM, no EA, Post-LASSO	52.1	48	30.8	9	113
(4) MCM, one-step EA, adaptive LASSO	114.3	112	40	47	187
(5) MCM, two-step EA, adaptive LASSO	46.8	41	28.9	2	111
(6) MCM, no EA, adaptive LASSO	84.8	87	36.5	13	156
(7) MOM, Post-LASSO	36.9	31	24.1	3	96
(8) MOM, adaptive LASSO	41.9	37	26.8	6	109
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	25.7	24	11.9	5	51
(10) MCM, one-step EA, LASSO	145.4	150	24.9	74	190
(11) Procedure (1) + additional confounders 1	21.9	19	14.1	0	60
(12) Procedure (11) + additional confounders 2	17.5	16	11.6	1	39
Cumulated employment during first 12 months					
(1) MCM, one-step EA, Post-LASSO	27.7	22	23.7	2	107
(2) MCM, two-step EA, Post-LASSO	3.3	2	3.8	0	16
(3) MCM, no EA, Post-LASSO	51.7	37	37.5	8	168
(4) MCM, one-step EA, adaptive LASSO	100	97	41	28	186
(5) MCM, two-step EA, adaptive LASSO	26.7	19	26.6	1	114
(6) MCM, no EA, adaptive LASSO	70.2	70	19.7	25	106
(7) MOM, Post-LASSO	34.5	30	18.8	9	100
(8) MOM, adaptive LASSO	52.6	47	29.2	13	121
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	13.2	11	9.7	0	47
(10) MCM, one-step EA, LASSO	107.1	102	19.5	70	150
(11) Procedure (1) + additional confounders 1	9.1	7	11	0	56
(12) Procedure (11) + additional confounders 2	6.3	6	5.3	0	19

Table continues on next page >

Table F.6 continued.

	Mean	Median	S.D.	Min	Max
	(1)	(2)	(3)	(4)	(5)
Cumulated employment during first 31 months					
(1) MCM, one-step EA, Post-LASSO	12.9	12	9.6	0	34
(2) MCM, two-step EA, Post-LASSO	3.1	2	3.8	0	14
(3) MCM, no EA, Post-LASSO	51.5	42	43	10	218
(4) MCM, one-step EA, adaptive LASSO	86.1	81	35.8	35	149
(5) MCM, two-step EA, adaptive LASSO	21.7	17	19.6	0	69
(6) MCM, no EA, adaptive LASSO	74	70	27.2	45	167
(7) MOM, Post-LASSO	35.6	29	22.6	12	107
(8) MOM, adaptive LASSO	72.3	66	32.3	26	144
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	7.9	4	14.4	0	75
(10) MCM, one-step EA, LASSO	73.6	79	19.4	0	101
(11) Procedure (1) + additional confounders 1	2.8	2	3.9	0	17
(12) Procedure (11) + additional confounders 2	3.6	2	5.6	0	27
Cumulated employment during months 25-31					
(1) MCM, one-step EA, Post-LASSO	4.8	2	6.8	0	29
(2) MCM, two-step EA, Post-LASSO	2.5	1	3.8	0	16
(3) MCM, no EA, Post-LASSO	40.4	27	36.5	10	160
(4) MCM, one-step EA, adaptive LASSO	78.6	72	36	22	176
(5) MCM, two-step EA, adaptive LASSO	13.7	11	12.8	1	46
(6) MCM, no EA, adaptive LASSO	61.4	59	19	35	137
(7) MOM, Post-LASSO	32.7	32	17.1	9	69
(8) MOM, adaptive LASSO	79.4	75	39.2	13	177
(9) MCM, one-step EA, Post-LASSO with radius-matching weights	1	0	1.7	0	6
(10) MCM, one-step EA, LASSO	45.4	52	38.4	0	99
(11) Procedure (1) + additional confounders 1	2.2	1	3.7	0	17
(12) Procedure (11) + additional confounders 2	1.9	0	3.6	0	17

Note: Description of the number of selected heterogeneity variables in all 30 sample splits for all considered implementations and outcomes. Standard deviations are abbreviated with S.D. in column (3). See Online Appendix F for more details about the different procedures. In Online Appendix F.5, we describe how we select additional confounders for procedures (11) and (12).