ORIGINAL ARTICLE

How sensitive are matching estimates of active labor market policy effects to typically unobserved confounders?

Stefan Tübbicke[1*]

Abstract

Using a rich and unique combined administrative-survey dataset, this paper explores how sensitive propensity score (PS) matching estimates of Active Labor Market Policies (ALMPs) based on the selection-on-observables assumption are to typically unobserved covariates. Using a sample of German unemployed welfare recipients, the analysis shows that typically unobserved factors such job search behavior, concessions willing to make for a job as well as (mental) health are in fact relevant confounders. However, results also show that matching on the PS using only typically observed covariates reduces imbalance in terms of typically unobserved covariates by about 46 percent in this setting. In line with this finding, the inclusion of typically unobserved covariates yields very similar estimates to estimates based on a standard specification. Hence, a standard matching approach based on rich and high quality administrative data appears to be sufficient to obtain estimates that are rather robust to unobserved confounding.

Keywords Propensity score matching, Observational studies, Selection bias, Active labor market policy, Evaluation

JEL Classification C21, D04, J68

1 Introduction

Matching and weighting based on the propensity score (PS, Rosenbaum and Rubin 1983) have become important tools for researchers aiming to flexibly estimate causal effects of some treatment using observational data. These methods have been widely applied in numerous fields of study such as Economics, Business, Sociology and Medicine. Most analyses use matching and weighting under an exogeneity assumption, assuming that controlling for observed background characteristics is sufficient to render the treatment as good as randomly assigned and remove bias from estimates. Comparisons of non-experimental with experimental estimates of treatment effects show, however, that such estimates may be plagued by

"hidden bias" (Rosenbaum 2002) due to unobserved confounders (see Heckman et al. 1997; Dehejia and Wahba 1999; Smith and Todd 2005, for examples). This is especially likely if the data are not rich enough, i.e. they do not contain a lot of background characteristics – especially pre-treatment outcomes – to condition on.

A prime example where matching and weighting estimators are routinely applied is the evaluation of active labor market programs (ALMPs) for the unemployed (see Caliendo and Künn 2011; Fitzenberger and Völter 2007; Harrer et al. 2020; Lechner and Wunsch 2009; Lechner et al. 2011, for examples). These studies are typically based on very detailed high-quality administrative data. Along with standard socio-demographics, household and regional characteristics, they contain daily information individuals' entire (un-)emploment, unemployment benefit receipt and ALMP history. Lechner and Wunsch (2013) use these rich administrative data and assess how sensitive effect estimates are to the omission of blocks of (observed) variables. They show, for example, that

*Correspondence:
Stefan Tübbicke
stefan.tuebbicke@iab.de
[1] Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany

information on individuals' health and characteristics of the last employer play an important role as conditioning variables. However, they cannot assess whether estimates are sensitive to unobserved factors.

This paper delivers insights in this regard and shows how sensitive matching and weighting estimates of ALMP effects are to *typically unobserved confounders* using a unique linked survey-administrative dataset from Germany. In addition to the high-quality administrative information, the data provides measures of typically unobserved variables relating to *attitudes towards work, job search behavior, willingness to make concessions for a new job, satisfaction in different domains, social participation, status and networks, (mental) health* and some *inter-generational* information. Numerous studies have shown that these factors are predictive of job-finding rates or exit rates from unemployment, making them prime candidates for potentially omitted but relevant confounders in the estimation of treatment effects. Based on a sample of unemployed welfare recipients in Germany, the paper investigates the importance of these typically unobserved variables for outcomes, selection into treatment, covariate balance as well as effect estimates.

This paper is closely-related to Caliendo et al. (2017) who perform a similar analysis using a sample of unemployment benefit (UB) recipients. As UB are only paid to individuals who have worked a minimum of 12 out of the 30 months before becoming unemployed, UB recipients typically have a better employment history, shorter unemployment duration and are more homogenous in general compared to unemployed welfare recipients. Hence, the analysis of this paper is expected to yield a stronger test regarding the role of typically unobserved variables as individuals are likely to be more heterogeneous not just in terms of typically observed confounders but potentially also in terms of typically unobserved characteristics. For example, Schubert et al. (2013) show that in 2006, welfare recipients were about 31 percent more likely to have a diagnosed mental illness compared to UB recipients. Since then, the prevalence of mental illnesses among welfare recipients has increased from roughly 33 to 45 percent as of 2021 (Deutscher Verein für öffentliche und private Fürsorge e.V. 2022). It is due to such unobserved heterogeneity that one may expect larger potential biases among welfare than among UB recipients if relevant but typically unobserved confounders are omitted from the estimation procedure. Moreover, this study differs from Caliendo et al. (2017) in the availability of typically unobserved covariates. On the one hand, this study has additional information on individuals' *attitudes towards work, willingness to make concessions for a new*

job, satisfaction in more domains than just life satisfaction as well as potentially crucial information on individuals' *(mental) health*. On the other hand, measures of *personality traits* and *expected ALMP participation probabilities* are not included in the survey (often enough) in order to be used in the present study.

In the context of welfare recipients, this paper shows that, overall, the typically unobserved variables observed through the survey data indeed are relevant confounders. Moreover, the results indicate that matching participants and comparison individuals on a standard estimate of propensity score based solely on typically observed covariates reduces imbalance in these typically unobserved covariates by roughly 46 percent. In line with this finding, differences between estimates of treatment effects using a standard specification and an extended specification that includes the typically unobserved covariates are relatively small and insignificant. Moreover, policy conclusions do not crucially depend on the availability of those typically unobserved confounders. Thus, it seems that – at least in the context considered – a rich specification based on typically observed confounders including pre-treatment outcomes may be sufficient to obtain reasonable estimates of treatment effects and to draw policy conclusions.

The remainder of this paper is organized as follows. Section 2 reviews identification and estimation of ALMP effects as well as the consequences of omitting relevant unobserved confounders from the analysis. Section 3 provides information on the institutional setting, the data used and shows some descriptive statistics for the sample. Section 4 performs the empirical analysis and Sect. 5 concludes.

2 Treatment effects and unobserved confounders

Using the potential outcomes framework by Roy (1951) and Rubin (1974), studies typically aim to estimate the average treatment effect on the treated (ATT)

$$\Delta^{ATT} = E[Y_i^1|D = 1] - E[Y_i^0|D_i = 1], \qquad (1)$$

where Y_i^1 refers to the outcome that is observed if person i received the treatment of interest, Y_i^0 is the outcome without treatment and D_i is a treatment indicator, taking on the value of one if person i received the treatment, zero otherwise. As Y_i^0 is unobservable for treated individuals, the second term in Eq. (1) has to be estimated from data on untreated persons.

Estimators based on the PS, defined as the conditional probability of receiving the treatment $Pr(D_i = 1|X_i)$, essentially re-weight non-participants to achieve balance

in terms of observed characteristics X_i. If treatment is assigned based on observed characteristics only, then this approach delivers unbiased estimates of treatment effects. Different versions of this underlying identification assumption have been termed unconfoundedness (Rosenbaum and Rubin 1983), selection-on-observables (Heckman and Robb 1985), or conditional-independence assumption (Lechner 2001). However, if treatment is also assigned based on unobserved characteristics U_i and these characteristics also affect the outcome of interest, then estimators based on the PS will be biased. The size of the bias after adjusting for the PS depends on the degree of imbalance U_i left as well as the strength of the association between U_i and Y_i. By construction, this bis cannot be directly estimated in a given study. However, one may inspect how influential an unobserved confounder must be to overturn the study's conclusions (Ichino et al. 2008; Oster 2019; Rosenbaum 2002).

Information on which typically unobserved confounders may be relevant in the case of ALMP evaluation can be gathered from empirical studies on the determinants of job finding rates. First, individuals' attitudes towards work are important in shaping their employment prospects. For example, individuals who view employed work as more central in their life display higher re-employment chances (Kanfer et al. 2001). Moreover, Zahradnik et al. (2016) show that unemployed individuals with a more intrinsic work motivation are less likely to be sanctioned, providing indirect evidence that they are more compliant regarding their obligation to cooperate with caseworkers. Second, a large body of literature shows that job search behavior is highly predictive of re-employment probabilities. Important factors include the intensity and the focus of job search as well as reservation wages (Altmann et al. 2018; Arni and Schiprowski 2019; Böheim et al. 2011; Krueger and Mueller 2016; Lichter and Schiprowski 2021; Koen et al. 2010). Third, the likelihood with which unemployed individuals find a job depends on whether they are willing to make concessions for a new job and if so, which ones (Andersson 2015; Caliendo et al. 2016; Christoph and Lietzmann 2022; Korpi and Levin 2001; Lietzmann et al. 2017). For example, greater geographical mobility is associated with better labor market outcomes (Yankow 2003). Fourth, subjective well-being is not only affected by unemployment (McKee-Ryan et al. 2005), but it also predicts how likely it is for individuals to get re-employed (Rose and Stavrova 2019). Sixth, social participation and networks have been shown to have a significant impact on labor market success (Bayer et al. 2008; Montgomery 1991). Seventh, general health as well as mental health are important determinants of individuals' employment chances (Butterworth et al. 2012; García-Gómez et al.

2013; Lötters et al. 2013; Schuring et al. 2007). Lastly, parental characteristics may constitute important omitted confounders as it is well documented that parental (un-)employment is predictive of later-in-life (un-)employment of their offspring (Fradkin et al. 2019; Pepper 2000).

All in all, these different findings highlight that administrative data, albeit rich, may be insufficient to obtain reasonable estimates of ALMP effects. Thus, it is imperative to assess the importance of these typically unobserved confounders on resulting effect estimates and policy conclusions drawn from evaluation studies based on the selection-on-observables assumption.

3 Institutional setting, data and descriptives

3.1 Institutional setting

There are two types of unemployment benefits (UB) in Germany. UB I are an insurance benefit. Individuals are eligible to receive UB I if they have contributed to the insurance system for at least 12 months out of the last three years when becoming unemployed. For individuals without children, the replacement rate is 60%, parents receive 67% of their last net salary. The maximum duration one can receive UB I is age-dependent. Individuals under 50 can receive UB I for up to 12 months, individuals 58 or older can receive UB I for up to two years.

The second type of benefits–UB II or simply welfare–is a means-tested flat-rate tax-financed benefit. To be eligible a person has to be able to work at least three hours a day and their household income must fall short of the legally defined social minimum. Hence, individuals can hold a job or even receive UB I and still be eligible for welfare provided that their household income is sufficiently low.

These differences in entry conditions result in very different populations of UB I and UB II recipients. While UB I recipients tend to have a relatively stable labor market history and relatively high re-employment chances, the labor market history of welfare recipients is often sparser and even if they are employed, earnings tend to be lower. Moreover, welfare recipients tend to stay in the system much longer. Official statistics by the Federal Employment Agency (2021) show that, at the end of 2021, about 66% of welfare recipients have been receiving said benefit for four years or longer. Thus, welfare recipients face stronger (and possibly unobserved) employment impediments compared to UB I recipients.

ALMPs are available to both UB I and welfare recipients. In fact, welfare recipients are can receive all kinds of ALMPs available to UB I recipients as well as some other measures designed exclusively for them. As activating unemployed welfare recipients is a key policy goal,

participation in ALMPs is often enforced using sanctions (Van den Berg and Vikström 2014; Van den Berg et al. 2022). The four most important types of ALMPs for welfare recipients are short-term training programs by external service providers, in-firm training, long-term training and One-Euro-Jobs.

Short-term training may for example be a job application training, a foreign language course or a training in a specific skill such as welding. In-firm training is essentially an unpaid internship. Long-term training programs may for example include remedial schooling for high-school dropouts to obtain a diploma as well as management, accounting or programming courses. Under certain conditions, they may even result in a vocational degree if completed. One-Euro-Jobs are a public employment creation program of additional jobs, i.e. non-market jobs, allowing jobseekers to earn one to two Euro per hour in addition to receiving welfare benefits.

Together, these programs made up around 80 percent of all ALMP spells among welfare recipients during our sample period. For the main analysis, the effect of any ALMP participation is estimated. As effects and selection patterns may be quite different across types of programs, heterogenous effects are also estimated for the four program types already mentioned as well as a remainder category, encompassing all other programs available to jobseekers on welfare.

3.2 Data and sample

This study uses the PASS-ADIAB dataset (Antoni et al. 2017), which combines administrative data from the Statistics department of the Federal Employment Agency with survey data for a representative sample of the German population. In addition to standard socio-demographic characteristics and household information, the administrative data provide daily information on individual's (un-)employment, benefit receipt and ALMP participation. Information on individuals' partners as well as their sanction history was merged from other administrative data sources.[1] Starting in 2007, the PASS ("Panel Arbeitsmarkt und soziale Sicherung", dubbed Panel Labor Market and Social Security) contains information on roughly 14,000 interviewees every year.[2] About half of interviewees are welfare recipients and their household members. The PASS contains a lot of additional information on issues such as attitudes towards work, job search behavior, concessions willing to make for a job, satisfaction in different domains, social participation and

networks, (mental) health as well as inter-generational transmission. For more information, see Trappmann et al. (2019).

For the analysis, this paper pools information on interviewees who are unemployed and receive welfare at the time of the interview from waves 5 (2011) to 8 (2014).[3] On the one hand, one may wish to use as many waves as possible to increase power of the statistical analysis. On the other hand, using additional waves tends to reduce the number of typically unobserved covariates which can be used in the analysis as not all questions are being asked in every wave. Hence, the choice of using waves 5 to 8 represents an attempt to balance these two objectives.

This approach yields a sample of 5819 individuals, 1009 of whom had an ALMP spell within four months after the interview and thus, are classified as participants. Non-participants are assigned a random hypothetical entry month in this four month window (Lechner 2002). Outcomes, namely regular employment (i.e. unsubsidized employment subject to social security contributions) as well as real monthly labor earnings are measured up to 36 months after (hypothetical) entry into treatment.

3.3 Descriptives

Panel A of Table 1 displays some descriptive statistics on typically observed covariates. First, one can see that participants are significantly younger on average compared to non-participants. Their mean age is roughly 42 years compared to 44 years among non-participants. Moreover, the share of females is significantly smaller among participants (42 percent) than non-participants (51 percent). Regarding place of residence, participants are more likely to live in Eastern Germany relative to non-participants. The share of individuals with a university degree is not statistically different between participants and non-participants. Lastly, panel A also provides information on the mean days spend in unemployment in the last 5 years in each sample. One can see that participants have spent, on average, 84 days less in unemployment than non-participants. Non-participants spent 982 days (or roughly 54 percent) of the last 5 years in unemployment. These results indicate that participants are somewhat positively-selected based on their labor market history and thus most likely also regarding their future employment prospects.

Panel B of Table 1 provides selected descriptive statistics for typically unobserved covariates. Regarding job

---

[1] The data source is called " Leistungshistorie Grundsicherung", which may be dubbed as welfare receipt history.

[2] Interviews are mostly conducted between February and September each year.

[3] In the sample of welfare recipients, roughly 80% of respondents are willing to have their survey data merged to the administrative data. Of those, about a third receive welfare and are registered as unemployed at the interview and thus, of interest for this study.

Table 1 Selected descriptives on covariates and outcomes

	ALMP		*p*-value
	Part.	Non-part.	
A. Typically observed covariates X			
Mean age in years	42.1	44.1	0.000
Share female	0.416	0.508	0.000
Share in Eastern Germany	0.395	0.336	0.001
Share with a university degree	0.045	0.055	0.180
Cumulated days in unemployment in last 5 years	898	982	0.000
B. Typically unobserved covariates U			
Share who searched for a job in last 4 weeks	0.574	0.432	0.000
Mean monthly net reservation wage in Euro	964	931	0.058
Share (mostly) willing to accept a long commute	0.416	0.310	0.000
Mean standardized life satisfaction	0.148	−0.031	0.000
Mean standardized subjective health	0.189	−0.040	0.000
Share of mothers with a university degree	0.074	0.053	0.015
C. Outcomes after 36 months			
Share in regular employment	0.295	0.211	0.000
Real monthly labor earnings	526	358	0.000
Number of observations	1,009	4,810	

This table shows sample means and shares of selected covariates and outcomes. *p*-values for differences between participants and non-participants are based on *t*-tests of equal means

search behavior, the Table shows the share of individuals who actively searched for a job in the last 4 weeks. While about 57 percent of participants searched for a job in the last month, only 43 percent of non-participants did so. Similarly, participants also show significantly higher mean reservation wages compared to non-participants. Moreover, about 42 percent of participants are (mostly) willing to accept a long commute in order to find a new job, among non-participants only 31 percent would be willing to do that. Regarding life satisfaction, participants' mean is roughly 15 percent of a standard-deviation above the sample mean, whereas non-participants' mean is three percent below the mean. Similar, but even more pronounced, differences are found in relation to individuals' subjective health. Lastly, participants are more likely to grow up with a university-educated mother than non-participants. All these differences are statistically significant at least at the 10 percent level and point towards positive selection based on typically unobserved covariates. For descriptives on the full set of typically unobserved variables, see Table 5 in the Appendix.

Lastly, Panel C of Table 1 also shows descriptives on outcomes used to evaluate the ALMPs. The comparison shows that participants perform better in the labor market after 36 months, both in terms of regular employment as well as earnings, than non-participants. While studies based on the selection-on-observables approach suggest causal effects of the same direction as this naive unconditional comparison, the question remains whether these findings are robust to the inclusion of typically unobserved covariates in the analysis when estimating effects.

4 Empirical analysis

This Section estimates causal effects of ALMPs on participants' labor market outcomes using two specifications. Similar to many evaluation studies based on observational data, the first ("standard") specification adjusts outcome differences between participants and non-participants for a large set of covariates that are observed through the administrative data. These include socio-demographics, household characteristics, partner characteristics, detailed labor market, benefit receipt and ALMP participation history as well as regional labor market controls.[4] The second ("extended") specification adjust outcome differences also for typically unobserved covariates obtained from the survey data. After

---

[4] Controlling for ALMP participation history is necessary in the context of welfare recipients as it is often the case that individuals participate in multiple programs throughout their unemployment history. This makes the interpretation somewhat difficult as past ALMP spells may interact with the effects we try to estimate. Due to the small sample, however, it is impossible to perform sub-sample analyses to inspect whether these interactions drive the results.

Table 2 Relevance of typically unobserved confounders

	Baseline Specification	Auxiliary Specifications							Extended Specification
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcomes of non-participants									
Regular employment									
R^2	0.1811	0.1823	0.1889	0.1862	0.1879	0.1818	0.1896	0.1822	0.2014
p-value of joint significance of U		0.1573	0.0000	0.0002	0.0000	0.4080	0.0000	0.4180	0.0000
Real monthly labor earnings									
R^2	0.1790	0.1806	0.1910	0.1851	0.1862	0.1807	0.1886	0.1803	0.2057
p-value of joint significance of U		0.0589	0.0000	0.0000	0.0000	0.0481	0.0000	0.3155	0.0000
Treatment assignment									
Pseudo-R^2	0.1003	0.1009	0.1059	0.1042	0.1027	0.1007	0.1037	0.1015	0.1124
p-value of joint significance of U		0.5120	0.0001	0.0039	0.0051	0.6735	0.0014	0.3579	0.0021
Control variables									
Typically observed covariates X	✓	✓	✓	✓	✓	✓	✓	✓	✓
Typically unobserved covariates U									
Attitudes towards work		✓							✓
Job search behavior			✓						✓
Concessions willing to make for a job				✓					✓
Satisfaction in different domains					✓				✓
Participation, social status and networks						✓			✓
(Mental) Health							✓		✓
Inter-generational information								✓	✓

This table compares summary statistics from OLS and logit regressions to inspect the relevance of (each block of) typically unobserved covariates

briefly reviewing kernel matching on the PS, this Section inspects the relevance of the typically unobserved confounders, examines balance of typically (un-)observed confounders before and after matching and compares effect estimates based on the standard and the extended specification. Effects are estimated for the pooled ALMP treatment indicator as well as for the main program types described in Section 3.1.

4.1 Estimation procedure using kernel matching

The analysis uses kernel matching on the PS, a widely-used technique to estimate causal effects under selection-on-observables. The estimation procedure is as follows: After having estimated the PS using a logit regression, common support in terms of the PS is inspected. As Heckman et al. (1998) show that lack of support can be a major source of evaluation bias, individuals outside of the common support are discarded from the analysis. This is done by removing participants from the estimation samples with values of the PS outside the range of non-participants (the so-called min-max criterion, see

Dehejia and Wahba 1999). Participants on support are then matched to non-participants based on the estimated PS. Using the popular Epanechnikov kernel, kernel matching places a larger weight on individuals that are closer in terms of the PS than individuals further away and avoids bad matches by discarding individuals outside of the user-chosen bandwidth (Caliendo and Kopeinig 2008). For simplicity, a standard bandwidth of 0.06 is used in the analysis. If balance is found to be sufficient after matching as in the main analysis, estimates of the ATT are the obtained as mean outcome differences in the matched sample. If imbalances remain after matching as in the program heterogeneity analysis, a linear regression with a treatment dummy and covariates is used on the matched sample to obtain estimates of the ATT. In any case, standard errors are estimated using the bootstrap with 999 replications (Bodory et al. 2020; MacKinnon 2006). Statistical inference is based on the normal approximation.

4.2 Relevance of the typically unobserved variables

This sub-section inspects the relevance of typically unobserved variables regarding the outcome and the assignment process. Relevance is tested block-wise as well as overall. Column one of Table 2 presents regression R^2 for OLS regressions of the outcomes as well as pseudo-R^2 from a logit regression on covariates using the standard specification. Columns two to eight individually add blocks of typically unobserved covariates and test for their joint significance using an F-test. This allows to asses which blocks of typically unobserved covariates have a significant association with the outcomes and the treatment assignment, controlling for the information already contained within the typically observed covariates. Lastly, column nine presents results for the extended specification, enabling a joint test regarding all typically unobserved covariates and their overall relevance for outcomes and selection into treatment. Results from the F-tests are presented using p-values of joint significance.

OLS regressions of the regular employment indicator as well as real monthly labor earnings after 36 months yield a regression R^2 of about 18 percent when using only covariates from the standard specification. Adding blocks of typically unobserved covariates one by one, we can see that especially job search behavior and (mental) health increase the R^2 to roughly 19 percent, closely followed by satisfaction in different domains and concessions willing to make for a job. Attitudes towards work only predict earnings, inter-generational information does not have a significant association with the outcomes. Adding all typically unobserved covariates increases regression R^2 to slightly over 20 percent for both outcomes. The joint tests of relevance show that these variables significantly predict outcomes on any traditional significance level.

Regressing the treatment indicator on the set of covariates included in the standard specification using a logit regression yields a pseudo-R^2 of roughly 10 percent. Adding the blocks one by one on top of the covariates from the standard specification yields pseudo-R^2s from 10.1 to 10.6 percent. The strongest increases in the pseudo-R^2 are achieved – in descending order – by adding job search related variables, covariates on (mental) health, concessions willing to make for a job and satisfaction in different domains. All of these blocks of typically unobserved variables significantly predict treatment assignment. Variables related to attitudes towards work, participation, social status and networks as well as inter-generational information are found to be insignificantly related to treatment. Adding all typically unobserved covariates in the extended specification yields a pseudo-R^2 of roughly 11.2 percent. Moreover, the joint F-test on all typically

unobserved covariates shows that these variables significantly predict treatment on any common significance level. The consequences of switching from the standard to the extended specification in terms of PS distribution can be inspected via kernel-density estimates in Fig. 2 in the Appendix, showing a shift of the distribution to the right and to the left for participants and non-participants, respectively.

Overall, the results show that the typically unobserved confounders provide additional information not contained in the typically observed confounders and thus, omitting them from the set of control variables may induce bias in treatment effect estimates of ALMPs.

4.3 Balancing quality

Next, the degree of covariate balance before and after matching in terms of observed and typically unobserved confounders is compared across specifications. To measure covariate balance, this paper follows the great majority of studies implementing PS-based estimators and uses the standardized (absolute) bias (Rosenbaum and Rubin 1983). The SB takes the absolute difference in means or sample shares for each covariate and standardizes it using the average standard deviation before matching[5]. Thus, using the SB it is possible to compare balance across variables that are measured on different scales.

Instead of reporting balancing for each covariate separately, Table 3 shows the mean SB (MSB) for blocks of typically (un-)observed variables. As expected, matching on the PS estimated using the standard specification, the MSB for all typically observed covariates X included in the specification is drastically reduced. Indeed, balance in terms of X can be regarded as excellent so that no additional regression-adjustment is necessary.

In this context, however, it is of greater interest to see how balancing of typically unobserved covariates changes when matching on typically observed covariates only. A reduction in the MSB for typically unobserved covariates can be seen as indication that standard PS specifications already capture (at least some) information that is included in these variables.

Indeed, after matching on the PS based on the standard specification, balancing regarding typically unobserved covariates U improves also. The MSB for U decreases from 10.2 to 5.5 percent, corresponding to a reduction in imbalance in terms of typically unobserved covariates by roughly 46 percent compared to before matching. Looking at blocks of typically unobserved covariates,

---

[5] To be exact, the SB for covariate k is $SB_k = 100 \cdot |\bar{X}_{k1} - \bar{X}_{k0}| / \sqrt{0.5 \cdot S_{k1}^2 + 0.5 \cdot S_{k0}^2}$, where $\bar{X}_{kD}$ is the covariate mean or sample share in treatment group D (before or after matching) and S_{kD} is the sample standard-deviation before matching.

Table 3 Mean (Absolute) standardized bias before and after matching

	Before matching	After	
		Standard specification	Extended specification
Typically observed characteristics X			
Socio-demographics	5.8	0.6	0.7
Household characteristics	4.1	0.6	0.4
Partner characteristics	2.2	0.6	0.6
Regional characteristics	11.0	0.5	0.3
Labor market history	10.3	1.1	0.7
Total (X)	8.7	0.9	0.7
Typically unobserved characteristics U			
ttitudes towards work	3.6	3.3	0.9
Job search behavior	14.9	8.5	0.5
Concessions willing to make for a job	17.5	7.8	0.4
Satisfaction in different domains	6.7	4.2	0.9
Participation, social status and networks	3.6	2.1	1.1
(Mental) Health	13.3	7.0	0.4
Inter-generational information	5.0	2.9	0.2
Total (U)	10.2	5.5	0.6
Total ($X + U$)	9.0	1.7	0.7
Control variables			
Typically observed covariates X		✓	✓
Typically unobserved covariates U			✓
Number of participants (on support)	1009	1009	1004

This table compares covariate balance across different blocks of covariates. Balance is measured using the mean (absolute) standardized bias (MSB, Rosenbaum and Rubin 1983)

the reduction in MSB is remarkably similar to the overall reduction in imbalance.

Comparing balancing results on typically unobserved covariates to Caliendo et al. (2017), it becomes evident that in the context of ALMP participation among unemployed welfare recipients, achieving balance in terms of typically observed covariates X reduces balance in terms of typically unobserved covariates U to a greater extent than among unemployment benefits (UB) recipients. Why might this be the case? As noted earlier, welfare recipients are expected to be a more heterogeneous group than UB recipients and treatment assignment may be more selective. Comparing pseudo-R^2 from the PS estimations, it becomes evident that typically observed covariates X do have more explanatory power regarding the treatment decision among welfare than UB recipients.

While Caliendo et al. (2017) report a pseudo-R^2 up to 8.7 percent, the pseudo-R^2 in this study is roughly 10 percent using the standard specification.[6] Hence, a larger degree of predictiveness of typically observed covariates regarding treatment may be helpful in reducing confounding due to typically unobserved covariates when evaluating ALMPs. However, these differences may also be driven by discrepancies regarding the sets of available typically (un-)observed covariates.[7]

4.4 Effect estimates

Having documented that balancing samples regarding typically observed covariates tends to also reduce imbalance in terms of typically unobserved covariates, it is interesting to inspect how much of a difference the

---

[6] Comparing pseudo-R^2 for the same programs yields even larger differences.

[7] For example, the standard specification of Caliendo et al. (2017) includes a total of 71 typically observed covariates, while this study uses 160 control variables in the standard specification.
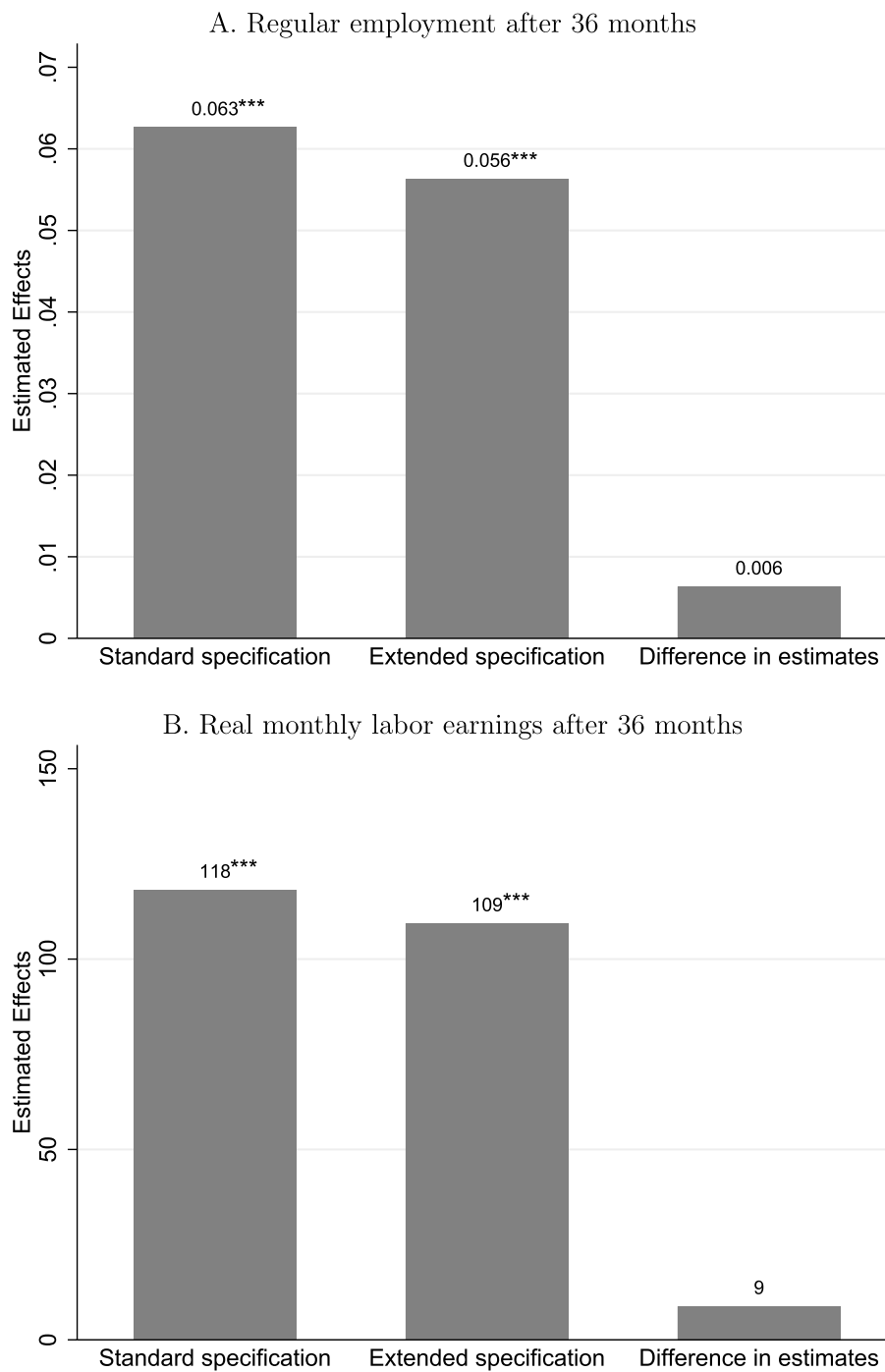
A. Regular employment after 36 months

B. Real monthly labor earnings after 36 months

Fig. 1 Main Results. This figure shows estimated ATTs. Statistical significance using bootstrapped standard errors on the 10/5/1% level is indicated by */**/***

inclusion of typically unobserved covariates in the estimation of the PS actually makes for the resulting treatment effects and policy conclusions. Figures 1 shows estimated treatment effects using kernel matching, both

for the standard as well as the extended specification. Moreover, the difference between both estimates is given and tested for statistical significance.

Based on the standard specification, estimates suggest that ALMP participation increases the chance of being

Table 4 Heterogeneity by type of ALMP

	Regular employment			Real monthly earnings		
	After 36 months			After 36 months		
	Standard specification	Extended specification	Difference in estimates	Standard specification	Extended specification	Difference in estimates
Short-term training	0.030	0.025	0.004	75*	70	5
	(0.024)	(0.025)	(0.008)	(42)	(43)	(15)
In-firm training	0.116***	0.105***	0.011	253***	250***	3
	(0.047)	(0.049)	(0.027)	(86)	(89)	(49)
Long-term training	0.141***	0.110***	0.031	285***	217***	68
	(0.052)	(0.052)	(0.030)	(105)	(108)	59.106
One-Euro-Jobs	− 0.041	− 0.034	− 0.007	− 34	− 19	− 15
	(0.025)	(0.024)	(0.012)	(43)	(43)	(20)
Other ALMPs	0.103***	0.100***	0.003	153***	133**	20
	(0.034)	(0.034)	(0.014)	(58)	(56)	(25)
Control variables						
Typically observed covariates	✓	✓		✓	✓	
Typically unobserved covariates		✓			✓	

This table shows estimated ATTs. Standard errors are obtained via bootstrapping with 999 replications. Statistical significance based on the normal approximation on the 10/5/1% level is indicated by */**/***. The kernel matching with additional regression adjustment employed uses a standard bandwidth of 0.06 (Leuven and Sianesi 2003)

in regular employment 36 months after starting treatment by 6.3 percentage points. Similarly, participants' real monthly labor earnings are expected to increase by 118 Euro after 36 months using the same specification. Switching to the extended specification yields effect estimates of 5.6 percentage points and 109 Euro for the employment and earnings outcomes, respectively. These differences of 0.6 percentage points and 9 Euro between estimates based on the standard and the extended specification are relatively small and not statistically significant on any common level. Moreover, even if differences were significant, including the typically unobserved confounders in the analysis would not alter conclusions about the effectiveness of ALMPs. Sensitivity checks show that these results are robust to using alternative estimation approaches (see Table 6) as well as alternative extended specifications (see Table 7).[8]

4.5 Program heterogeneity

As effects of and selection into different ALMPs can be quite heterogeneous, this section briefly re-estimates effects for different kinds of programs, namely short-term training, in-firm training, long-term training, One-Euro-Jobs as well as "other" programs, entailing all other programs available to unemployed welfare recipients during the study period. This leads to relatively small samples compared to the main analysis.[9] Results can be found in Table 4.

Focusing on results based on the standard specification first, estimated effects in-firm, long-term training programs as well as "other" programs imply substantial positive effects on employment and earnings after 36 months. Estimated effects for short-term training are also positive, but smaller and statistically insignificant, most likely due sample size restrictions. Regarding One-Euro-Jobs,

---

[8] Two alternate extended specifications are used. First, job search variables are dropped due to potential anticipation issues (van den Berg et al. 2009). Second, inter-generational variables are dropped due to their insignificance in the treatment and in the outcome equations. Both extended specifications yield essentially the same estimates.

[9] Numbers of participants are 352 (short-term training), 127 (in-firm training), 108 (long-term training), 227 (One-Euro-Jobs) and 195 ("other"). Insufficient balancing quality after matching, especially for in-firm with an MSB of 7 percent and long-term training with an MSB of 5.7 percent, required additional regression adjustment to control for bias due to residual confounding (Caliendo and Kopeinig 2008).

point estimates are negative and statistically insignificant. Overall, these results closely resemble main findings by other evaluation studies which estimate causal effects based on administrative data only (see Bernhard and Kruppe 2012; Harrer et al. 2020; Harrer and Stockinger 2022; Huber et al. 2011, for examples). Next, we compare these estimates to estimates obtained using the extended specification. For training programs, estimates of employment effects are smaller by 0.4 percentage points for short-term training, 1.1 percentage points for in-firm training and 3.1 percentage points for long-term training. Estimates for One-Euro-Jobs also decrease in magnitude and become closer to zero, but remain negative and statistically insignificant. In all cases, differences in estimated employment effects are highly statistically insignificant. Estimated effects on earnings follow a similar pattern with differences in estimates being relatively small and highly statistically insignificant. Hence, the inclusion of typically unobserved confounders in the estimation of the PS does not yield different estimated effects or policy conclusions compared to relying on typically observed confounders only.

5 Discussion and conclusion

Using a unique combined administrative-survey dataset, this paper inspects whether the evaluation of ALMPs for unemployed welfare recipients in Germany is robust to the inclusion of typically unobserved covariates in the analysis. While the usually unobserved factors analyzed are significant predictors of treatment the outcomes of interest, differences in estimated effects between a standard specification relying on covariates typically observed in administrative datasets and the extended specification are relatively small and statistically insignificant. This supports findings by Caliendo et al. (2017) who perform a similar analysis for a sample of unemployment benefit recipients.

Moreover, the inspection of covariate balance reveals that, in this context, aiming to achieve balance in terms of typically observed covariates also reduces imbalance in terms of typically unobserved covariates by about 46 percent. This reinforces the notion that matching on rich data – especially pre-treatment outcomes – also helps reducing bias due to unobserved confounders. A plausible explanation of this phenomenon is that pre-treatment outcomes may already have been affected by those unobserved confounders in the past and thus, conditioning on pre-treatment outcomes may help proxy for unobserved factors.

In comparison to Caliendo et al. (2017), the reduction in imbalance achieved in terms of typically unobserved covariates by conditioning on typically observed covariates only is relatively large. At the same time, measures of predictiveness from logit regressions of the treatment indicator on typically observed covariates suggest stronger selection into treatment among welfare than unemployment benefit recipients. Hence, it appears that a more predictive set of covariates may be helpful in reducing potential biases due to typically unobserved factors. However, differences in the degree of imbalance reduction may also be driven by different sets of typically (un-)observed covariates. One could only try to disentangle these factors if one were to have access to both datasets by comparing results for different sets of typically (un-)observed control variables. However, such an analysis is beyond the scope of this paper.

Overall, the results indicate that estimated effects of ALMPs and resulting policy conclusions are robust to the inclusion of typically unobserved confounders in the analysis. Hence, rich administrative data seem to be sufficient to obtain reasonable estimates of causal effects in this context. Nonetheless, one should not over-interpret the findings of this paper. Although the analysis uses numerous typically unobserved covariates, effect estimates may still be sensitive to other factors not observed through the survey data. Moreover, it is uncertain whether these results generalize to the evaluation of ALMPs in other countries, for example due to differences in institutional features. Furthermore, it may be that estimates of causal effects of other kinds of treatments, for example in the medical context, based on the selection-on-observable assumption are more prone to bias due to unobserved confounding. These issues remain uncertain and require additional research in the future.

Appendix

Tables 5, 6 and 7 and Fig. 2

Table 5 Descriptives on all typically unobserved covariates

	ALMP		*p*-value
	Part.	Non-part.	
Attitudes towards work			
Share (mostly) agreeing:			
Work is just a means to make money	0.549	0.565	0.388
Having a job is the most important in life	0.640	0.679	0.019
A job makes you feel included	0.885	0.885	1.000
I would also work if I did not need the money	0.751	0.737	0.371
Job search behavior			
Searched for a job in last 4 weeks	0.574	0.432	0.000
Searched for a specific job in last 4 weeks	0.128	0.129	0.928
Searched for different kinds of jobs in last 4 weeks	0.238	0.178	0.000
Searched for any job in last 4 weeks	0.128	0.129	0.928
Number of search channels used	2.642	2.104	0.000
Had a job interview in last 4 weeks	0.220	0.160	0.000
Monthly net reservation wage in Euro	964	931	0.058
Types of concessions willing to make for a job			
Share (mostly) willing:			
Willing to accept a long commute	0.416	0.310	0.000
Willing to accept a low wage	0.298	0.247	0.001
Willing to work long hours	0.405	0.318	0.000
Willing to work over-qualified	0.548	0.426	0.000
Willing to accept unfavorable working conditions	0.438	0.324	0.000
Willing to move for a job	0.180	0.134	0.000
Willing to do long-distance relationship	0.184	0.149	0.008
Mean standardized degree of satisfaction in different domains			
With life	0.148	−0.031	0.000
With standard of living	-0.015	0.003	0.618
Wth apartment	0.000	−0.000	0.996
Participation, social status and networks			
Mean standardized subjective social participation	− 0.005	0.001	0.854
Mean standardized subjective position on the social ladder	− 0.030	0.006	0.291
Share active in a sports/music/cultural club	0.164	0.139	0.058
Mean number of close friends/family outside of own HH	5.248	5.141	0.401
(Mental) Health			
Mean standardized subjective health	0.189	−0.040	0.000
Mean standardized mental health	0.058	−0.012	0.046
Mean no. of doctors visits in last 3 months	2.574	3.336	0.000
Share admitted to a hospital in last year	0.148	0.179	0.011
Intergenerational information			
Share of mothers with a vocational deg	0.297	0.317	0.240
Share of mothers with a university degree	0.074	0.053	0.015
Share of fathers with a vocational degree	0.151	0.174	0.068
Share of fathers with a unversity degree	0.100	0.084	0.149
Share of fathers employed when respondent was 15	0.772	0.783	0.462
Share of fathers self-employed when respondent was 15	0.075	0.081	0.516

This table compares covariate means or shares across treatment groups. *p*-values are obtained from equality of means tests

Table 6 Estimated effects using other matching and weighting approaches

	Regular employment after 36 months			Real monthly earnings after 36 months		
	Std.	Ext.	Diff.	Std.	Ext.	Diff.
Kernel matching (baseline)	0.063***	0.056	0.006	118***	109***	9
	(0.017)	(0.018)	(0.006)	(29)	(29)	(9)
Caliper radius matching	0.054***	0.058***	− 0.005	107***	113***	− 3
	(0.019)	(0.019)	(0.014)	(32)	(32)	(18)
Local linear matching	0.066***	0.060***	0.006	124***	116***	8
	(0.018)	(0.018)	(0.006)	(29)	(30)	(9)
Inverse probability weighting	0.063***	0.057***	0.006	117***	111***	7
	(0.017)	(0.017)	(0.004)	(28)	(29)	(7)
Doubly robust	0.058***	0.054***	0.005	110***	104***	6
	(0.017)	(0.017)	(0.004)	(28)	(28)	(7)
Control variables						
Typically observed covariates	✓	✓		✓	✓	
Typically unobserved covariates		✓			✓	

This table shows estimated ATTs. Standard errors are obtained via bootstrapping with 999 replications. Statistical significance based on the normal approximatio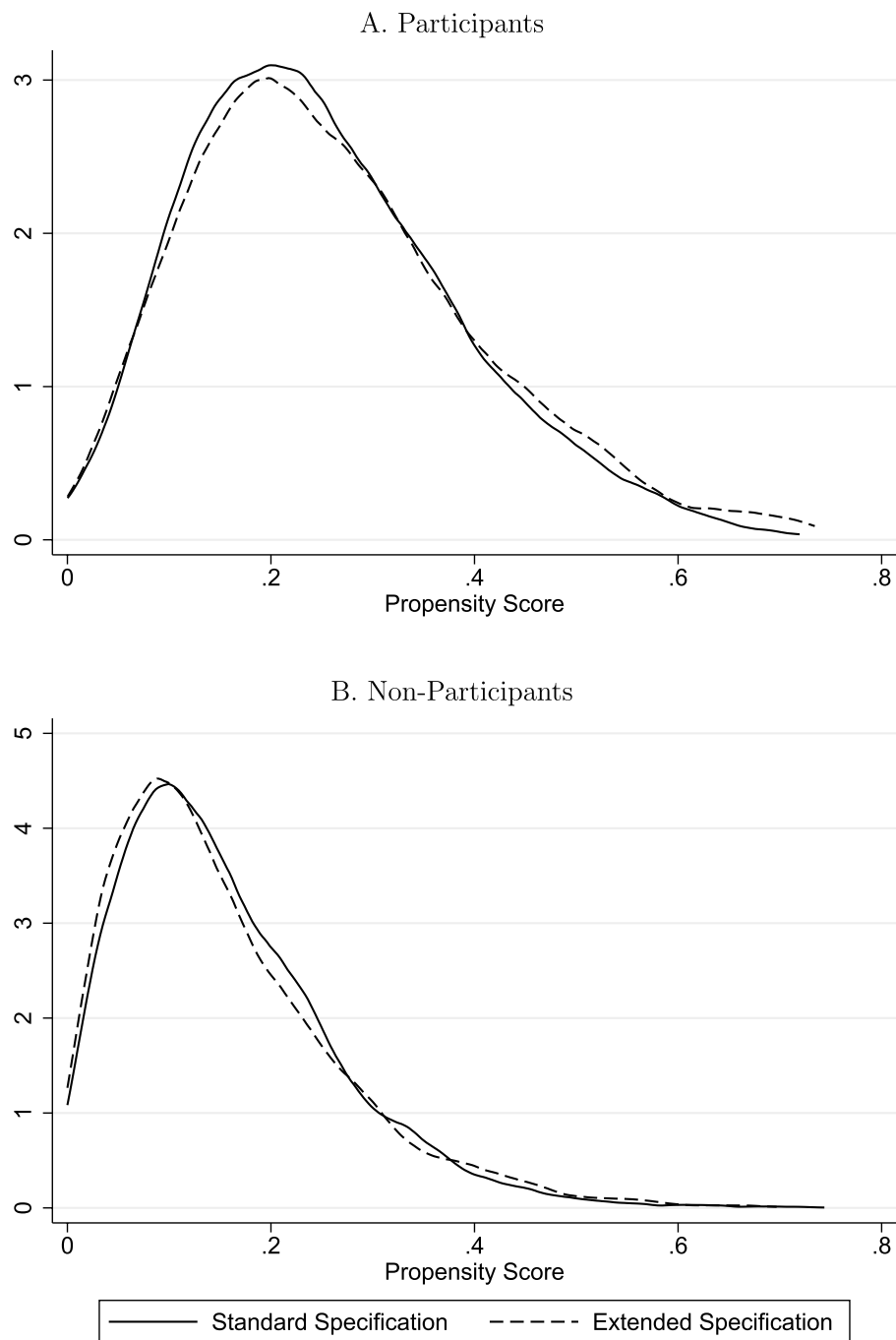n on the 10/5/1% level is indicated by * / ** / ***. Local linear matching use a standard bandwidth of 0.06. The caliper for the radius matching is set equal to three times the 90% quantile of the distribution of differences in the propensity score from a nearest neighbor matching with replacement (Huber et al. 2015). The doubly robust estimator combines inverse probability weighting with parametric outcome adjustment (Robins et al. 1995)

Table 7 Estimated Effects using other Extended Specifications

	Regular employment after 36 months			Real monthly earnings after 36 months		
	Std.	Ext.	Diff.	Std.	Ext.	Diff.
No job search variables	0.063***	0.056***	0.007	118***	105***	13
	(0.016)	(0.016)	(0.005)	(29)	(29)	(8)
No inter-generational info	0.063***	0.057***	0.005	118***	110***	8
	(0.016)	(0.016)	(0.005)	(29)	(29)	(9)
Control variables						
Typically observed covariates	✓	✓		✓	✓	
Typically unobserved covariates		✓			✓	

This table shows estimated ATTs. Standard errors are obtained via bootstrapping with 999 replications. Statistical significance based on the normal approximation on the 10/5/1% level is indicated by * / ** / ***

A. Participants

B. Non-Participants

Fig. 2 Propensity Score Distribution. This figure shows kernel-density estimates of the propensity score distribution for the standard and the extended specification

Author contributions
Does not apply.

Declarations

References

Altmann, S., Falk, A., Jäger, S., Zimmermann, F.: Learning about job search: a field experiment with job seekers in Germany. J. Pub. Econ. **164**, 33–49 (2018)

Andersson, K.: Predictors of re-employment: a question of attitude, behavior, or gender? Scand. J. Psychol. **56**(4), 438–446 (2015)

Antoni, M., Dummert, S., Trenkle, S., et al.: PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975–2015. FDZ Datenrep. **6**, 2017 (2017)

Arni, P., Schiprowski, A.: Job search requirements, effort provision and labor market outcomes. J. Pub. Econ. **169**, 65–88 (2019)

Bayer, P., Ross, S.L., Topa, G.: Place of work and place of residence: informal hiring networks and labor market outcomes. J. Political Econ. **116**(6), 1150–1196 (2008)

Bernhard, S., Kruppe, T.: Effectiveness of further vocational training in Germany—empirical findings for persons receiving means-tested unemployment benefits. J. Appl. Soc. Sci. Stud. **132**(4), 501–526 (2012)

Bodory, H., Camponovo, L., Huber, M., Lechner, M.: The finite sample performance of inference methods for propensity score matching and weighting estimators. J. Bus. Econ. Stat. **38**(1), 183–200 (2020)

Böheim, R., Horvath, G.T., Winter-Ebmer, R.: Great expectations: past wages and unemployment durations. Labour Econ. **18**(6), 778–785 (2011)

Butterworth, P., Leach, L.S., Pirkis, J., Kelaher, M.: Poor mental health influences risk and duration of unemployment: a prospective study. Soc Psychiatry Psychiatri. Epidemiol. **47**(6), 1013–1021 (2012)

Caliendo, M., Kopeinig, S.: Some practical guidance for the implementation of propensity score matching. J. Econ. Surv. **22**(1), 31–72 (2008)

Caliendo, M., Künn, S.: Start-up subsidies for the unemployed: long-term evidence and effect heterogeneity. J. Pub. Econ. **95**(3–4), 311–331 (2011)

Caliendo, M., Künn, S., Uhlendorff, A.: Earnings exemptions for unemployed workers: the relationship between marginal employment, unemployment duration and job quality. Labour Econ. **42**, 177–193 (2016)

Caliendo, M., Mahlstedt, R., Mitnik, O.A.: Unobservable, but unimportant? the relevance of usually unobserved variables for the evaluation of labor market policies. Labour Econ. **46**, 14–25 (2017)

Christoph, B., Lietzmann, T.: The relevance of job-related concessions for unemployment duration among recipients of means-tested benefits in Germany. J. Soc. Policy **51**(2), 242–267 (2022)

Dehejia, R.H., Wahba, S.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J. Am. Stat. Assoc. **94**(448), 1053–1062 (1999)

Deutscher Verein für öffentliche und private Fürsorge e.V.: Empfehlungen des Deutschen Vereins für öffentliche und private Fürsorge e.V. zur Unterstützung von Personen mit psychischen Beeinträchtigungen und psychischen Erkrankungen in der Grundsicherung für Arbeitsuchende (SGB II). Tech. rep., Soziale Sicherungssysteme und Sozialrecht (2022)

Federal Employment Agency: Langzeitleistungsbezieher - Deutschland, West/Ost, Länder und Jobcenter (Monatszahlen). (2021)

Fitzenberger, B., Völter, R.: Long-run effects of training programs for the unemployed in East Germany. Labour Econ. **14**(4), 730–755 (2007)

Fradkin, A., Panier, F., Tojerow, I.: Blame the parents? How parental unemployment affects labor supply and job quality for young adults. J. Labor Econ. **37**(1), 35–100 (2019)

García-Gómez, P., Van Kippersluis, H., O'Donnell, O., Van Doorslaer, E.: Long-term and spillover effects of health shocks on employment and income. J. Hum. Res. **48**(4), 873–909 (2013)

Harrer, T., Stockinger, B.: First step and last resort: one-euro-jobs after the reform. J. Soc. Policy **51**(2), 412–434 (2022)

Harrer, T., Moczall, A., Wolff, J.: Free, free, set them free? Are programmes effective that allow job centres considerable freedom to choose the exact design? Int. J. Soc. Welf. **29**(2), 154–167 (2020)

Heckman, J.J., Robb, R.: Alternative methods for evaluating the impact of interventions: an overview. J. Econom. **30**(1–2), 239–267 (1985)

Heckman, J.J., Ichimura, H., Todd, P.E.: Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econ. Stud. **64**(4), 605–654 (1997)

Heckman, J., Ichimura, H., Smith, J., Todd, P.: Characterizing selection bias using experimental data. Econometrica **66**(5), 1017–1098 (1998)

Huber, M., Lechner, M., Wunsch, C., Walter, T.: Do german welfare-to-work programmes reduce welfare dependency and increase employment? Ger. Econ. Rev. **12**(2), 182–204 (2011)

Huber, M., Lechner, M., Steinmayr, A.: Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour. Empir. Econ. **49**(1), 1–31 (2015)

Ichino, A., Mealli, F., Nannicini, T.: From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? J. Appl Econom **23**(3), 305–327 (2008)

Kanfer, R., Wanberg, C.R., Kantrowitz, T.M.: Job search and employment: a personality-motivational analysis and meta-analytic review. J. Appl. Psychol **86**(5), 837 (2001)

Koen, J., Klehe, U.-C., Van Vianen, A.E., Zikic, J., Nauta, A.: Job-search strategies and reemployment quality: the impact of career adaptability. J. Vocat Behav. **77**(1), 126–139 (2010)

Korpi, T., Levin, H.: Precarious footing: temporary employment as a stepping stone out of unemployment in Sweden. Work Employ. Soc. **15**(1), 127–148 (2001)

Krueger, A.B., Mueller, A.I.: A contribution to the empirics of reservation wages. Am. Econ. J. Econ. Policy **8**(1), 142–79 (2016)

Lechner, M.: Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. Econometric evaluation of labour market policies, pp. 43–58. Springer, Berlin (2001)

Lechner, M.: Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. J. Royal Stat. Soc. Series A **165**(1), 59–82 (2002)

Lechner, M., Wunsch, C.: Are training programs more effective when unemployment is high? J. Labor Econ. **27**(4), 653–692 (2009)

Lechner, M., Wunsch, C.: Sensitivity of matching-based program evaluations to the availability of control variables. Labour Econ. **21**, 111–121 (2013)

Lechner, M., Miquel, R., Wunsch, C.: Long-run effects of public sector sponsored training in West Germany. J. Eur. Econ. Assoc. **9**(4), 742–784 (2011)

Leuven, E., Sianesi, B.: PSMATCH2: stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical software components. Boston College Department of Economics, Newton (2003)

Lichter, A., Schiprowski, A.: Benefit duration, job search behavior and re-employment. J. Pub. Econ. **193**, 104326 (2021)

Lietzmann, T., Schmelzer, P., Wiemers, J.: Marginal employment for welfare recipients: stepping stone or obstacle? Labour **31**(4), 394–414 (2017)

Lötters, F., Carlier, B., Bakker, B., Borgers, N., Schuring, M., Burdorf, A.: The influence of perceived health on labour participation among long term unemployed. J. Occup. Rehabilit. **23**(2), 300–308 (2013)

MacKinnon, J.G.: Bootstrap methods in Econometrics. Econ. Record **82**, S2–S18 (2006)

McKee-Ryan, F., Song, Z., Wanberg, C.R., Kinicki, A.J.: Psychological and physical well-being during unemployment: a meta-analytic study. J. Appl. Psychol. **90**(1), 53 (2005)

Montgomery, J.D.: Social networks and labor-market outcomes: toward an economic analysis. Am. Econ. Rev. **81**(5), 1408–1418 (1991)

Oster, E.: Unobservable selection and coefficient stability: theory and evidence. J. Bus. Econ. Stat. **37**(2), 187–204 (2019)

Pepper, J.V.: The intergenerational transmission of welfare receipt: a nonparametric bounds analysis. Rev. Econ. Stat. **82**(3), 472–488 (2000)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am. Stat. Assoc. **90**(429), 106–121 (1995)

Rose, D., Stavrova, O.: Does life satisfaction predict reemployment? Evidence form German panel data. J. Econ. Psychol. **72**, 1–11 (2019)

Rosenbaum, P.R.: Overt bias in observational studies. Observational studies, pp. 71–104. Springer, Berlin (2002)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)

Roy, A.D.: Some thoughts on the distribution of earnings. Oxf. Econ. Pap. **3**(2), 135–146 (1951)

Rubin, D.: Estimating causal effects to treatments in randomised and nonrandomised studies. J. Educ. Psychol. **66**(5), 688–701 (1974)

Schubert, M., Parthier, K., Kupka, P., Krüger, U., Holke, J., Fuchs, P.: Menschen mit psychischen Störungen im SGB II. Tech. rep, IAB-Forschungsbericht (2013)

Schuring, M., Burdorf, L., Kunst, A., Mackenbach, J.: The effects of ill health on entering and maintaining paid employment: evidence in European countries. J. Epidemiol. Commun. Health **61**(7), 597–604 (2007)

Smith, J.A., Todd, P.E.: Does matching overcome lalonde's critique of nonexperimental estimators? J. Econom. **125**(1–2), 305–353 (2005)

Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., Schwarz, S., Teichler, N., Unger, S., Wenzig, C.: Data resource profile: panel study labour market and social security (PASS). Int. J. Epidemiol. **48**(5), 1411–1411g (2019)

Van den Berg, G.J., Vikström, J.: Monitoring job offer decisions, punishments, exit to work, and job quality. Scand. J. Econ. **116**(2), 284–334 (2014)

van den Berg, G.J., Bergemann, A.H., Caliendo, M.: The effect of active labor market programs on not-yet treated unemployed individuals. J. Eur. Econ. Assoc. **7**(2–3), 606–616 (2009)

Van den Berg, G.J., Uhlendorff, A., Wolff, J.: The impact of sanctions for young welfare recipients on transitions to work and wages, and on dropping out. Economica **89**(353), 1–28 (2022)

Yankow, J.J.: Migration, job change, and wage growth: a new perspective on the pecuniary return to geographic mobility. J. Reg. Sci. **43**(3), 483–516 (2003)

Zahradnik, F., Schreyer, F., Moczall, A., Gschwind, L., Trappmann, M.: Wenig gebildet, viel sanktioniert? zur Selektivität von Sanktionen in der Grundsicherung des SGB II. Z. Soz. **62**(2), 141–180 (2016)

Publisher's Note