



# New evidence on long-term effects of start-up subsidies: matching estimates and their robustness

Marco Caliendo<sup>1,2,3,4</sup> · Stefan Tübbicke<sup>1</sup>

Received: 24 August 2018 / Accepted: 30 April 2019 / Published online: 21 May 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

The German start-up subsidy (SUS) program for the unemployed has recently undergone a major makeover, altering its institutional setup, adding an additional layer of selection and leading to ambiguous predictions of the program's effectiveness. Using propensity score matching (PSM) as our main empirical approach, we provide estimates of long-term effects of the post-reform subsidy on individual employment prospects and labor market earnings up to 40 months after entering the program. Our results suggest large and persistent long-term effects of the subsidy on employment probabilities and net earned income. These effects are larger than what was estimated for the pre-reform program. Extensive sensitivity analyses within the standard PSM framework reveal that the results are robust to different choices regarding the implementation of the weighting procedure and also with respect to deviations from the conditional independence assumption. As a further assessment of the results' sensitivity, we go beyond the standard selection-on-observables approach and employ an instrumental variable setup using regional variation in the likelihood of receiving treatment. Here, we exploit the fact that the reform increased the discretionary power of local employment agencies in allocating active labor market policy funds, allowing us to obtain a measure of local preferences for SUS as the program of choice. The results based on this approach give rise to similar estimates. Thus, our results indicating that

---

The authors thank Lutz Bellmann, two anonymous reviewers, the editor and participants at the 7th ifo Dresden Workshop on Labour Economics and Social Policy, the University of Barcelona's Workshop on Unemployment and Labor Market Policies, the 2017 conference of the European Society for Population Economics, the LISER workshop on Causal Inference, Program Evaluation, and External Validity, and the 2017 conference of the European Association of Labor Economists for helpful discussions and valuable comments. We are grateful to the Institute for Employment Research (IAB) for cooperation and institutional support within the research Project No. 1755.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00181-019-01701-9>) contains supplementary material, which is available to authorized users.

---

✉ Marco Caliendo  
caliendo@uni-potsdam.de

Extended author information available on the last page of the article

SUS are still an effective active labor market program after the reform do not appear to be driven by “hidden bias.”

**Keywords** Start-up subsidies · Policy reform · Matching · Instrumental variables

**JEL Classification** J68 · H43 · C14 · C26 · L26

## 1 Introduction

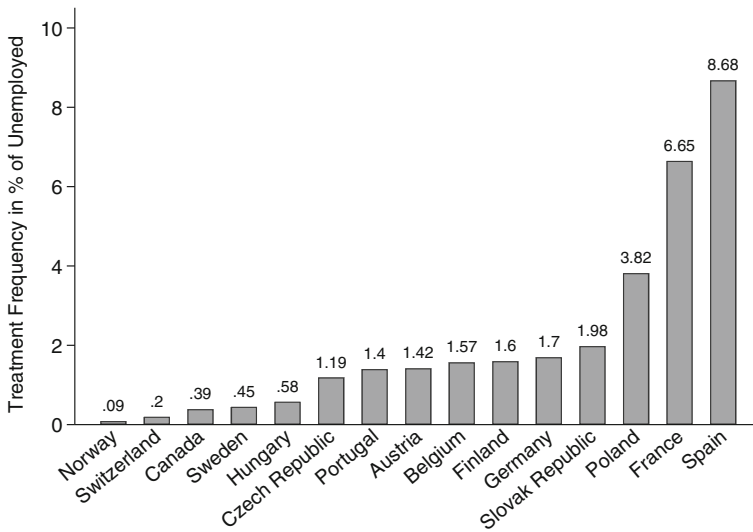
Start-up subsidies (SUS) for the unemployed are an unconventional active labor market program (ALMP). They help unemployed individuals to escape unemployment by incentivizing them to start their own business and securing their livelihood during the first uncertainty-ridden months of the start-up. The usage of SUS has recently been on the rise: According to official statistics by the OECD (2015), participation in this type of programs is high, whereby in Spain 8.7% of the stock of unemployed participated in a start-up incentive program, closely followed by France with 6.7% and Poland with 3.8%.<sup>1</sup> The empirical evidence on the effectiveness of SUS as an ALMP is more scarce compared to other programs such as training measures, although the body of evidence is growing. In general, almost all studies find positive and relatively large effects on individual labor market outcomes.<sup>2</sup> However, all of the mentioned studies rely on the conditional independence assumption (Lechner 2001), also known as the selection-on-observables assumption (Heckman and Robb 1985), whereby they assume that—conditional on a vector of observable characteristics—treatment is as good as randomly assigned. Thus, these estimates are susceptible to “hidden bias” if the researcher does not observe all relevant pre-treatment characteristics.

In this paper, we provide first evidence on long-term individual labor market effects of the German SUS program called “Gründungszuschuss”—which we dub new start-up subsidy (NSUS)—after its reform in 2011. The reform altered the institutional setup of the program and was mainly intended to reduce spending on SUS (see Bernhard and Grüttner 2015). The reduction in spending was achieved through abandoning entitlement to the program, thereby giving caseworkers at local employment agencies more discretionary power to reject applicants as well as by instituting large budget cuts of about €800m from 2011 to 2012. Additionally, monetary support to participants was reduced, leading to ambiguous predictions of the post-reform effectiveness of the program.<sup>3</sup> Furthermore, abolishing the entitlement to the program introduced an additional layer of selection, thus potentially reducing the credibility of making inference using methods relying on the conditional independence assumption. Therefore,

<sup>1</sup> For an overview of the importance of SUS programs in OECD countries, see Fig. 1.

<sup>2</sup> For example, effect estimates are provided by Tokila (2009) for Finland, Duhautois et al. (2015) for France, Caliendo and Künn (2011) and Wolff et al. (2016) for Germany, O’Leary (1999) for Hungary and Poland, Perry (2006) for New Zealand, Rodríguez-Planas and Jacob (2010) for Romania and Behrenz et al. (2016) for Sweden. An in-depth review of estimated effects and the institutional setup is given by Caliendo (2016).

<sup>3</sup> For a detailed description of the program before and after the 2011 reform for the NSUS in Germany, estimated short-term program effects and a discussion of the importance of the institutional setup of the program, see Bellmann et al. (2018).



**Fig. 1** Participants in SUS in OECD Countries 2015. *Source:* OECD (2015), own calculations

studying the effects of SUS and their sensitivity to deviations from the identifying assumption under this post-reform setting provides an interesting case study to shed some light on the reliability of estimates under these circumstances. This is especially true because SUS programs in other countries operate with a similar selection mechanism, requiring joint decision-making by the unemployed individual and the caseworker (see, e.g., Behrenz et al. 2016, on the current Swedish program). In addition, many countries' SUS programs are designed in a similar fashion where, support is granted by paying out a series of periodic transfers to recipients, mostly dependent on previous labor earnings (O'Leary 1999).

Our main approach to estimating long-term effects of the German NSUS makes use of propensity score matching (PSM), as introduced by Rosenbaum and Rubin (1983). Within the matching framework, we assess the robustness of our estimates with respect to implementation-related issues and deviations from the conditional independence assumption (CIA). Going beyond the standard matching approach, we also provide estimates using an instrumental variable (IV) identification approach based on regional variation in the likelihood of receiving treatment. Here we exploit the fact that the reform increased the discretionary power of local employment agencies in allocating ALMP funds, allowing us to obtain a measure of local preferences for the NSUS as the ALMP of choice. As a proxy for these preferences, we use regional application approval rates for the NSUS, conditional on local labor market conditions. Using a sample of 1248 participants and 1204 non-participants, our matching results indicate persistent and positive long-term effects on individual employment probabilities and labor earnings up to 40 months after entering the program. Our sensitivity analysis within the matching framework shows that these findings are robust with respect to both issues related to the implementation of the matching estimators as well as deviations from the CIA. Finally, our estimates based on the IV strategy also give rise to similar

estimates. Thus, our findings of large and positive effects of SUS for participants are unlikely to be driven by “hidden bias.”

The remainder of this paper is organized as follows. Section 2 provides an overview of the institutional details of the NSUS program, gives details on the selection by caseworkers and discusses theoretical predictions on the post-reform effectiveness. Section 3 describes our dataset and presents some descriptives. Section 4 discusses the necessary identifying assumptions of our matching approach. Section 5 provides our main estimates and discusses effect heterogeneity. Section 6 performs our extensive sensitivity analyses, and Sect. 7 concludes.

## 2 Institutional setup of the new start-up subsidy

**The post-reform program** In its current form, the NSUS has been in place since December 2011.<sup>4</sup> In order to be eligible for the program, unemployed individuals have to be entitled to at least another 150 days of unemployment benefits and obtain proof of sustainability for their business plan issued by an independent institution like the chamber of commerce. In contrast to previous programs, there is no legal entitlement to the subsidy under the reformed NSUS conditional on meeting the aforementioned eligibility criteria.<sup>5</sup> Thus, caseworkers at local employment agencies (LEAs) can deny access to the program to eligible applicants. Successful applicants receive a monthly payment equivalent to their unemployment benefits, which depends on previous labor earnings, plus a lump sum of €300 for the first 6 months after entering into the program. Participants may also apply for a second benefit period that only provides monthly payment of the lump sum for an additional 9 months. Thus, in total, the program provides financial support to participants for a maximum of 15 months. In our sample, about 57% of participants received transfers for the second benefit period. The average total support was €10,350 for participants.

**Selection by caseworkers** For the purpose of our analysis, it holds particular importance to understand the selection mechanism that determines participation and non-participation. Selection into different ALMPs is regulated by §7, social code book III, which states that caseworkers make an individual decision on the necessity of activation measures and the appropriateness of certain measures for the unemployed individual. When making this decision, the abilities of the unemployed individuals are to be taken into consideration. For the case of SUS, this means that the applicant needs to be considered as sufficiently entrepreneurial to run a business. Bernhard and Grüttner (2015) provide important qualitative evidence on caseworkers’ behavior and the way in which they and their LEAs handled the transition to more discretionary power induced by the reform of the program. In their interviews with stakeholders from different LEAs, they find that the most commonly cited reason why applica-

<sup>4</sup> It is currently the only SUS program available to unemployment benefits I recipients. Unemployment benefits II recipients, which are mostly long-term unemployed or individuals with very sparse employment history, are eligible for a different program called “Einstiegsgeld,” which is not the focus of this study.

<sup>5</sup> For a description and evaluations of the predecessor programs, see Caliendo and Künn (2011, 2014, 2015), Caliendo et al. (2016).

tions were rejected was a sufficiently large number of applicant-specific vacancies in the local labor market, as judged by the individual caseworker. This is consistent with the so-called placement priority as defined by §4, social code book III, which states that caseworkers are only meant to consider ALMPs as an option for unemployed individuals if they are necessary for the re-integration of the individual. Taken together, this information suggests that the most important confounders in our analysis are the individuals' re-employment probability in the absence of treatment and their entrepreneurial affinity. Arguably, the former can be controlled for relatively well using pre-treatment labor market outcomes, local labor market conditions and measures of human capital. However, the latter is generally unobservable and difficult to proxy for (see Caliendo et al. 2016, for a detailed discussion of this issue) and thus at the center of our sensitivity analysis in Sect. 6.

**Theoretical predictions** By comparison, the pre-reform program required fewer days of unemployment benefits to be eligible and the first benefit period lasted 9 months, instead of 6. Shortening of the first benefit period might lead to larger effects through a reduction in moral hazard, although it may also reduce the effectiveness of the program due to lower financial support to help overcome capital constraints. Moreover, the additional layer of selection induced by the reform may potentially lead to larger effects due to the previously mentioned "placement priority" by selecting individuals who benefit more from the program. Furthermore, effects may be different simply due to macroeconomic forces. Overall, these considerations lead to ambiguous predictions on the magnitude of effects after the reform relative to before.

### 3 Data and descriptive analysis

#### 3.1 Data

For our analysis, we use a random sample of previously unemployed participants who joined the program between February and June 2012. Data on participants from January were not used as most entrants still joined the program under pre-reform conditions, i.e., they applied before the reform was in place. Our comparison group consists of individuals who were unemployed for at least one day, eligible for the program but did not apply for it in this period. Both samples were drawn from the Integrated Labor Market Biographies (IEB) of the Federal Employment Agency (FEA). Our dataset combines extensive register data from the IEB with informative survey data collected via two computer-assisted telephone interviews around 20 and 40 months after entering the program. In order to reduce survey costs, non-participants to be interviewed were selected via a pre-matching strategy to avoid interviewing individuals with very dissimilar observed characteristics compared to actual participants. For this purpose, for each participant who entered the program in month  $m$ , 20 non-applicants were randomly drawn from the unemployed population and assigned month  $m$  as their month of fictitious entry. A nearest neighbor matching was conducted based on basic variables such as age, gender, education, regional labor market types and short-term labor market history as measured by the employment status at the end of 2011, the

timing of entry into unemployment as well as the (hypothetical) entry month. Aside from ensuring the basic comparability of participants and non-participants, this yields a balanced duration of the unemployment spell from the date of entry into unemployment and the (hypothetical) month of entry into the program across the two groups.<sup>6</sup> Among non-participants, only nearest neighbors were contacted for the survey.

Due to the combination of register and survey data, our dataset contains extensive covariates on individuals' labor market history, previous earnings, socio-demographics, human capital, ALMP history, participants' start-up characteristics, intergenerational information as well as usually unobserved personality traits. From the surveys, we are able to use labor market outcome data up to 40 months after (hypothetical) entry into the program. The final dataset contains 1248 participants and 1204 non-participants. Participants in our sample account for about 17% of all entrants into the SUS program during our sampling frame.<sup>7</sup>

### 3.2 Some descriptives

In this part, we provide a brief descriptive overview of our sample of participants and comparison individuals. Table 1 provides summary statistics on socio-demographics, human capital, labor market history, intergenerational transmission, regional macro-economic conditions and personality traits. For a more extensive overview of descriptive statistics on covariates, see Table A.1 in Appendix. Outcome statistics can be found in Table 2.

**Pre-treatment characteristics** Participants are on average about 42 years old and about one year younger than non-participants. In addition, participants are less likely to be female. While a sizable fraction of about 43% of participants have attained a general upper secondary school degree, which grants access to the German university system, only 28% of non-participants have such a degree. With respect to labor market history, participants spent on average 10% of the last 10 years in unemployment. On the other hand, non-participants were unemployed for 17% of the last 10 years. Short-term employment history shows that participants were employed for about 7.7 months in employment in the previous year before entering the program. On average, non-participants were employed for about one month less during this time period. The majority of participants and non-participants (67% and 52%, respectively) were in dependent employment before entering unemployment. While 5.4% of participants were self-employed before entering unemployment, only 1.2% of non-participants had the same employment status. An economically significant fraction of 35% of the treated and 25% of comparison individuals have at least one self-employed parent, which is described as one of the key drivers in the decision to become self-employed by the entrepreneurship literature (e.g., see Dunn and Holtz-Eakin 2000; Lindquist et al. 2016). Participants and non-participants also differ with respect to personality

<sup>6</sup> Participants spent on average 2.8 months in unemployment before entering the program. Our sample of non-participants was unemployed for 2.7 months on average prior to the assigned date of entry. The  $p$  value of a  $t$ -test of equality of means is about 0.22.

<sup>7</sup> According to the FEA, about 7400 individuals entered the program between February and June 2012.

**Table 1** Selected descriptives for control variables

	Part.	Non-part.	<i>p</i> value
<i>Socio-demographics</i>			
Age at (hypothetical) entry	42.77	43.94	0.002
Less than 25 years	0.012	0.006	0.103
25 to less than 35 years	0.206	0.192	0.383
35 to less than 45 years	0.336	0.322	0.478
45 to less than 56 years	0.195	0.191	0.817
56 years and older	0.252	0.289	0.037
Female	0.425	0.509	0.000
<i>Human capital</i>			
Highest schooling degree			
Lower secondary school	0.107	0.198	0.000
Middle secondary school	0.278	0.376	0.000
Upper secondary school (specialized)	0.170	0.128	0.004
Upper secondary school (general)	0.433	0.282	0.000
No schooling degree	0.013	0.016	0.537
<i>Labor market history</i>			
Fraction of time in unemployment in the last 10 years	0.101	0.170	0.000
One year before (hypothetical) entry			
Months employed	7.773	6.699	0.000
Months in labor market program	0.401	0.390	0.834
Employment status before entering unemployment			
Dependent employment	0.674	0.516	0.000
Self-employment/family worker	0.054	0.012	0.000
School/apprenticeship	0.017	0.020	0.567
Disable to work/unemployable	0.033	0.135	0.000
Other	0.081	0.263	0.000
Daily income from last employment (Euro)	81.39	61.60	0.000
Duration of last unemployment spell (months)	3.808	4.390	0.010
<i>Intergenerational information</i>			
Father and/or mother was born abroad	0.151	0.177	0.079
Father and/or mother is/was self-employed	0.349	0.252	0.000
Father employed when respondent 15 years old	0.913	0.855	0.000
<i>Geographic region</i>			
Northern Germany	0.196	0.167	0.059
Eastern Germany	0.333	0.374	0.033
Southern Germany	0.155	0.182	0.080
<i>Local macroeconomic conditions</i>			
Local unemployment rate in %	7.905	7.927	0.853
Ratio of vacancies to unemployed	16.92	17.24	0.439

**Table 1** continued

	Part.	Non-part.	<i>p</i> value
GDP per capita in 2011	31.26	30.67	0.088
Local start-up rate out of unemployment	0.058	0.058	0.932
Local self-employment rate	0.112	0.114	0.002
<i>Personality traits</i>			
Big Five			
Conscientiousness	4.383	4.106	0.000
Extraversion	3.692	3.297	0.000
Agreeableness	3.401	3.368	0.300
Neuroticism	2.365	2.677	0.000
Openness	3.723	3.255	0.000
Other personality traits			
Readiness to take risks	5.777	5.316	0.000
Locus of control	4.260	3.785	0.000
Patience	6.370	6.380	0.911
Impulsiveness	5.388	5.409	0.809
General self-efficacy	4.391	4.062	0.000
Number of observations	1248	1204	

Reported are sample shares for dummy variables and means for multi-valued variables. *p* values are based on *t* tests of equal means

traits. For example, participants are on average more conscientious, extraverted, open to new experience and more risk tolerant than comparison individuals.

This shows that although comparison individuals have been pre-matched and thus their sample is not representative of the underlying general unemployed population, there remain significant in-sample differences in key characteristics between the treated and non-treated.<sup>8</sup>

**Labor market outcomes** Table 2 provides mean labor market outcomes for participants and non-participants at 20 and 40 months after (hypothetical) entry. At the first interview—about 20 months after entry—88.8% of participants and only 3.7% of non-participants are self-employed. Despite being smaller at the second interview, the gap remains substantial. For our causal analysis later, we will focus on an overall employment indicator, without discriminating between self-employment and regular employment, as both types of employment are seen as a successful integration into the labor market. At the first interview—20 months after entry—95.8% of participants and 61.3% of non-participants are in self- or regular employment. At the second interview—after 40 months—the overall employment rate is slightly lower for participants with 93.3% and higher for non-participants with 67.4%. At both interviews, there is a substantial raw gap in net monthly labor earnings in favor of the participants.

<sup>8</sup> The significant gap between treated and comparison group characteristics is due to the fact that pre-matching was done in a very coarse way to ensure minimal overlap between the two groups.



**Table 2** Descriptives for outcome variables

	Part.	Non-part.	<i>p</i> value
<i>Labor market outcomes after 20 months</i>			
Main labor market status			
Self-employed	0.884	0.037	0.000
Self- or regular employed	0.958	0.615	0.000
Unemployed	0.021	0.208	0.000
Earnings			
Net monthly earnings (Euro)	1901.2	886.6	0.000
<i>Labor market outcomes after 40 months</i>			
Main labor market status			
Self-employed	0.804	0.038	0.000
Self- or regular employed	0.938	0.676	0.000
Unemployed	0.022	0.110	0.000
Earnings			
Net monthly earnings (Euro)	2264.4	1046.3	0.000
Number of observations <sup>a</sup>	1248	1204	

Reported are sample shares for labor market status variables and means for the earnings variables. *p* values are based on *t* tests of equal means

<sup>a</sup>The number of observations for the earnings variables is slightly lower due to item non-response

## 4 Main empirical approach

The goal of our causal analysis in this and the next section is to estimate the treatment effects of the SUS program on individuals' labor market outcomes in terms of overall employment and earned income. We rely on the well-known potential outcomes framework, mainly attributed to Roy (1951) and Rubin (1974). Our main focus is to estimate the average treatment effect on the treated (ATT)

$$\tau_{ATT} = E(Y^1 | D = 1) - E(Y^0 | D = 1), \tag{1}$$

where  $Y^1$  and  $Y^0$  are potential outcomes with and without treatment and  $D$  is a treatment indicator (= 1 if individual received a SUS). Since  $E(Y^0 | D = 1)$  is generally unobservable, it has to be inferred from data on non-participants' outcomes. However, simply using the mean outcome of non-participants will lead to biased estimates in the absence of random assignment of treatment due to differential characteristics between the two groups.

**Propensity score matching** PSM techniques aim to eliminate selection bias by balancing a rich set of observable characteristics  $X$  across the two groups. To give consistent estimates, the so-called CIA

$$Y^0 \perp\!\!\!\perp D | P(X) \tag{2}$$

needs to hold, where  $P(X) = Pr(D = 1 | X)$  is the propensity score. In addition, one has to assume overlap ( $P(X) < 1 \forall X$ ) and rule out spillover effects of treatment (*stable unit treatment value assumption*). As noted by Imbens and Wooldridge (2009), if these three assumptions hold, we can estimate (1) as the simple mean difference between treated and comparison individuals on the *re-weighted* sample as

$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_i^1 - \sum_{j=1}^{N_0} \hat{w}_j Y_j^0, \quad (3)$$

where  $N_1$  and  $N_0$  are the number of treated and untreated observations and  $i$  and  $j$  are their respective indices. Estimated balancing weights  $\hat{w}_j$  are obtained through matching, where the resulting weights satisfy  $\sum_{j=1}^{N_0} \hat{w}_j = 1$  and  $\hat{w}_j \geq 0$ .

**Inference** In order to account for the multi-step estimation procedure of PSM, we make use of re-sampling methods for hypothesis testing. In particular, we obtain  $p$  values by bootstrapping the  $t$ -statistic with 999 replications, as this has been shown to have better properties than bootstrapping standard errors directly (see Huber et al. 2015; MacKinnon 2006, for details).

**Risk of hidden bias** For the CIA to be a valid assumption,  $X$  must contain all such variables that simultaneously determine selection into treatment and the outcome of interest (Lechner and Wunsch 2013). Consequently, if there is some unobserved characteristic  $U$  that has an impact on treatment assignment and the outcome, the CIA will fail. Put formally,  $\hat{\tau}_{ATT} \xrightarrow{p} \tau_{ATT} + b$  with  $b \neq 0$  if the true treatment probability is given by

$$P_i := Pr(D = 1 | X = x_i, U = u_i) = F(x_i' \beta + \gamma u_i) \quad (4)$$

where  $\gamma \neq 0$  and  $E[Y^0 | X = x, U = u] \neq E[Y^0 | X = x, U = u']$  for  $u \neq u'$ . The size of the inconsistency  $b$  depends on the selectivity parameter  $\gamma$  and the responsiveness of  $Y^0$  with respect to  $U$ . Since the reform of the NSUS introduced an additional layer of selection,  $\gamma$  may be larger in magnitude and estimates more susceptible to “hidden bias” (see Rosenbaum 2002, for more details on the problem of “hidden bias”). Thus, careful sensitivity analyses are necessary.

#### 4.1 Specification and estimation of the propensity score

Our extensive dataset allows us to control for a wide range of pre-treatment characteristics. Our baseline propensity score specification includes variables containing information on socio-demographics such as age, gender, health status, German citizenship, marital status and single parent status, number of children and the presence of young children. Human capital attainment is included using the highest schooling degree, professional education and qualification. In order to break the dependence between  $D$  and  $Y^0$ , it is arguably most important to include a detailed and sufficiently

flexible specification of labor market and earnings history. We do this by adding information on short- and long-term unemployment history, short- to medium-term employment and treatment history, the employment status before unemployment, previous occupation, the size of unemployment benefits received as well as last labor earnings. We make the specification flexible by either including categorical dummies for important confounders or—as in the case of previous earnings—using data-driven selection of fractional polynomials (Sauerbrei and Royston 1999). In addition, we also include a battery of regional characteristics to control for different local labor markets. For this purpose, we include regional dummies as well as explicit control for local macroeconomic conditions and self-employment activity. The baseline specification also includes a number of interaction terms, which were added iteratively to improve subsequent matching quality (see next Sect. 4.2).

As mentioned in Sect. 2, another potentially important confounder is the entrepreneurial affinity of individuals. We attempt to proxy for it using some variables that are seen to be important factors in determining the decision to become self-employed, such as previous self-employment status, intergenerational transmission of self-employment as observed through our survey data and the regional controls on start-up activity out of unemployment and the share of self-employed in the general labor force. Aiming to strike a balance regarding the comparability of our results to most previous evidence for Germany as well as evidence for other countries, we make use of intergenerational information but abstain from including our measures of personality traits in our baseline specification. However, as part of our sensitivity analysis later on, we will extend this standard set of variables with some measures of personality traits or non-cognitive skills since they are likely to be correlated with entrepreneurial affinity. The baseline specification is estimated using a probit regression on the pooled sample, as a Chow test of different selection patterns into treatment for men and women could not be rejected. The details of the specification, estimated coefficients and results from the Chow test can be found in Table A.2. Figure 2 shows the resulting predicted values of the propensity score used to estimate balancing weights in the next step.

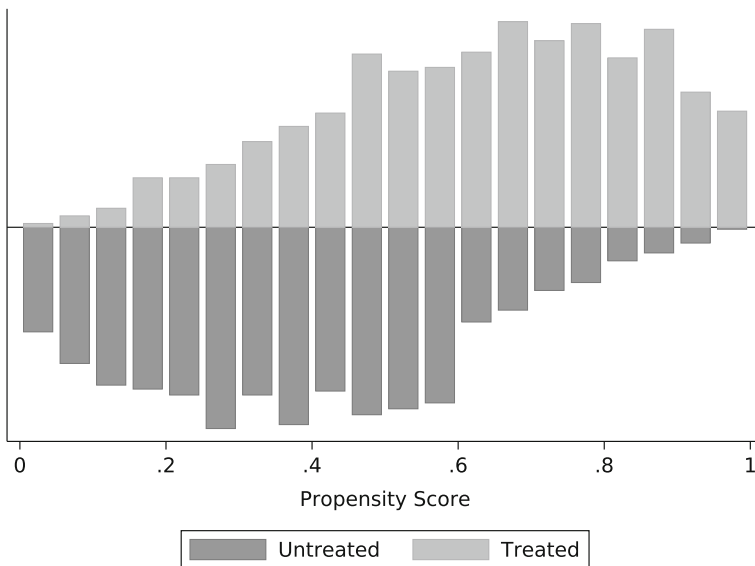
## 4.2 Matching to improve balance

In our baseline design, we use the estimated propensity score in combination with nonparametric kernel matching with an Epanechnikov kernel to estimate balancing weights.<sup>9</sup> The kernel bandwidth is chosen to maximize post-matching balance.<sup>10</sup> In order to avoid extrapolation, we impose common support by restricting the analysis to the subset of treated individuals who satisfy

$$\left\{ i \in S_1 : \hat{p}(x_i) < \max_j (\hat{p}(x_j) \mid D = 0) \right\}, \quad (5)$$

<sup>9</sup> The matching is performed using the *psmatch2* ado-package by Leuven and Sianesi (2003).

<sup>10</sup> In the spirit of Imai et al. (2008), a grid search is performed, choosing the bandwidth that maximizes balance by minimizing the pseudo-R<sup>2</sup> after matching. We found this to be the case for  $h = 0.13$ .



**Fig. 2** Propensity score distribution—Baseline specification. *Note* This graph shows the distribution of estimated propensity scores for the treated and comparison group using a probit regression based on the baseline specification including information on socio-demographics, human capital, labor market history, intergenerational transmission, and regional controls for local labor market conditions and self-employment activity. For details on the specification and estimated coefficients, see Table A.2

where  $S_1$  denotes the set of all treated units. Since matching on the propensity score does not control for differences in covariates directly, the appropriateness of the propensity score specification has to be judged against the resulting balancing quality (Rosenbaum and Rubin 1983). Table 3 provides several commonly used indicators for the balance achieved.

Kernel matching dramatically improves in-sample balance as measured by several indicators. There remain no significant mean differences in the matched sample at any traditional level using a  $t$  test of equal means. This is supported by a reduction in the mean absolute standardized bias from 11.5% to 2.3% through matching. Moreover, inspecting the distribution of absolute standardized biases reveals that the number of covariates with standardized biases with relatively large differences is drastically reduced. For example, the number of covariates with a standardized bias above 7% is reduced from 38 to zero. Similarly, the number of covariates with a bias of at least 5% but less than 7% decreased from 28 to just 8 in the matched sample. Following Sianesi (2004), pseudo- $R^2$  of the propensity score estimation decreases to 1.5% in the matched sample and the null hypothesis of all covariates having no predictive power regarding treatment status cannot be rejected at virtually any significance level. The balancing measures based on the propensity score due to Rubin (2001) also point toward a drastic increase in balancing quality. Rubin's  $B$ —defined as the standardized mean difference in the linear index ( $x\hat{\beta}$ ) of the propensity score—decreases from over 100% to 29.3%, while the ratio of the propensity score's variance in the treated and untreated sample (Rubin's  $R$ ) remains close to one. In addition, quantile–quantile plots for the important

**Table 3** Balancing quality

	Before matching	After matching
Number of variables with significant differences in means <sup>a</sup>		
At 1%-level	38	0
At 5%-level	50	0
At 10%-level	58	0
Number of variables with absolute standardized bias <sup>b</sup>		
0 to less than 1%	9	29
1 to less than 3%	9	35
3 to less than 5%	12	24
5 to less than 7%	28	8
more than 7%	38	0
Mean absolute standardized bias in %	11.47	2.26
(Re-)estimation of the propensity score <sup>c</sup>		
Pseudo- $R^2$	.204	.015
<p>value of joint-significance test</p>	0.000	0.999
Other measures		
Rubin's $B^d$	111.0	29.3
Rubin's $R^e$	1.00	1.37
Number of variables	96	96
Number of participants off support		58

Different indicators are shown for covariate balancing before and after Epanechnikov kernel matching using a bandwidth of 0.13

<sup>a</sup>The number of variables with significantly different means is based on a  $t$  test of equality of means

<sup>b</sup>The standardized absolute bias of a variable is the mean difference between treatment and control group as a percentage of the square root of the mean of pre-matched variances of both groups

<sup>c</sup>Following Sianesi (2004) Pseudo- $R^2$  and  $p$  value of joint significance from a probit estimation on the unmatched and the matched sample are also calculated

<sup>d</sup>Rubin's  $B$  is the standardized mean difference of the linear index of the propensity score of treatment and control group

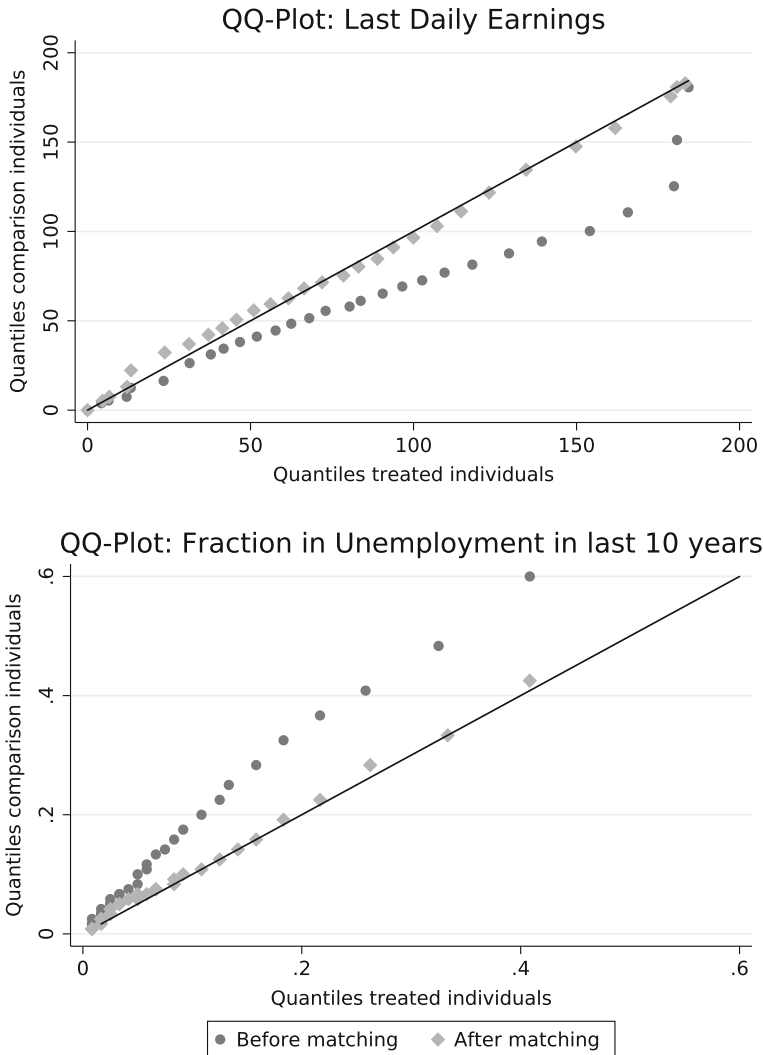
<sup>e</sup>Rubin's  $R$  is the variance ratio of the propensity score index of the treated to control sample

pre-treatment outcomes “fraction of time spent in unemployment in the last 10 years” and “last daily earnings” follow the 45-degree line quite closely, indicating successful balancing of the distribution of these important covariates after matching (see Fig. 3). Overall, balancing quality can be regarded as sufficient to proceed with the outcome analysis.

## 5 Estimates based on propensity score matching

### 5.1 Main estimates

Panel A in Table 4 presents the estimated average treatment effects for participants using our baseline empirical approach as described in Sect. 4.



**Fig. 3** Graphical analysis of balancing quality. *Note* This figure plots distribution quantiles for the treated against those of the untreated, both for the raw and the matched sample. Scatter dots following the 45 degree line indicate covariate balance for continuous variables

Consistent with the existing literature on SUS for the unemployed, we find persistent and large effects on both the employment probability and monthly net earned income of program participants. Participants are about 28 percentage points more likely to be in self- or regular employment, and they earn on average about €760 more than the matched comparison group at the first interview 20 months after entering the program. Regarding long-term effects, estimates suggest that participants are on average 21.5 percentage points more likely to be self-employed or regular employed 40 months after entry. Effects on net monthly earned income are even greater at the

**Table 4** Baseline results and sensitivity analyses with respect to implementation

	Outcomes after 20 months		Outcomes after 40 months	
	Self- or regular employment	Net monthly earned income	Self- or regular employment	Net monthly earned income
<i>A. Baseline results</i>				
	0.280*** (0.000)	764.2*** (0.000)	0.215*** (0.000)	980.2*** (0.000)
<i>B. Sensitivity with respect to choice of CDF <math>F(\cdot)</math></i>				
Logit	0.279*** (0.000)	762.6*** (0.000)	0.213*** (0.000)	977.3*** (0.000)
Robit	0.278*** (0.000)	787.6*** (0.000)	0.208*** (0.000)	979.7*** (0.000)
<i>C. Sensitivity with respect to common support imposition</i>				
$f(\hat{p}) > 0.2$	0.283*** (0.000)	739.6*** (0.000)	0.218*** (0.000)	948.2*** (0.000)
$[0; \alpha^*]$	0.281*** (0.000)	748.8*** (0.000)	0.216*** (0.000)	959.0*** (0.000)
<i>D. Sensitivity with respect matching/weighting algorithm</i>				
NN matching	0.285*** (0.000)	803.1*** (0.000)	0.239*** (0.000)	909.7*** (0.000)
IPW	0.283*** (0.000)	794.5*** (0.000)	0.216*** (0.000)	990.1*** (0.000)
LMW radius	0.283*** (0.000)	784.2*** (0.000)	0.234*** (0.000)	933.1*** (0.000)

The table shows ATT estimates using different PSM approaches.  $p$  values are shown in parentheses underneath the point estimate, obtained by bootstrapping the  $t$ -statistic using 999 replications (MacKinnon 2006; Huber et al. 2015). \*\*\*/\*\*/\* denote significance at the 1/5/10 % level.

Panel A shows our baseline results using kernel matching, common support imposition via min/max criterion using a probit regression. Panel B tests the sensitivity with respect to the parametric regression of the propensity score. Panel C tests whether the way in which we impose common support affects our results significantly, defining a minimum density  $f(\hat{p})$  in the comparison group (Heckman et al. 1997) or choosing an optimal common support interval  $[0; \alpha^*]$  (Crump et al. 2009). Finally, Panel D tests the robustness regarding different estimators:  $k = 1$  nearest neighbor matching, inverse probability weighting with standardized weights and radius matching with bias adjustment due to Lechner et al. (2011)

second interview compared to the first one, whereby participants gain around €980 by joining the program. These estimated effects are both statistically significant and economically substantial. The size of the effects is of similar magnitude to what Caliendo and Künn (2011) found for an older SUS program introduced by the “Hartz reforms” in 2003. Compared to estimates for the pre-reform program by Caliendo et al. (2016), our point estimates are around 11 percentage points larger with respect to employment effects and about €250 larger in terms of effects on earned income. Thus, our empirical results may be cautiously interpreted as pointing toward a positive role of the institutional changes regarding the program’s effectiveness despite their ambiguous



**Fig. 4** Effect heterogeneity regarding employment effects. *Note* This graph plots the estimated ATTs against the estimated counterfactual means  $\hat{E}[Y^0 | D = 1]$  for subsamples. Results are obtained by repeating the steps of the main analysis for each subsample separately. Sample splits are performed based on binary indicators regarding age (age  $\geq 45$  or not), education (= high if individual has a (specialized) higher secondary school degree), GDP per capita (= high if the individual lives in a region with above median GDP per capita) and gender

theoretical impact, indicating room for improvement of SUS programs by changing entry conditions and support.

## 5.2 Effect heterogeneity

In order to gain further insight into how effects vary with respect to certain pre-treatment characteristics and tease out potential channels through which the program works, we estimate ATTs for subgroups according to age, education, local GDP per capita and gender.<sup>11</sup> The results are displayed in Fig. 4.

In general, one can say that the groups that have a particularly low estimated counterfactual probability of being in self- or regular employment 40 months after entering the program display the largest estimated gains from participating. These are the lower educated without a (specialized) higher secondary schooling degree—which grants access to the university system—and workers who are at least 45 years old. Our findings support the view that low-skilled individuals benefit more from participating in SUS programs than high-skilled workers. Larger effects for older workers either point toward more entrepreneurial success among older founders or reflect more diffi-

<sup>11</sup> The entire estimation procedure is repeated for each subsample. Balancing indicators and propensity score distributions for the subsamples are available upon request from the authors. Generally, matching quality is somewhat worse due to smaller sample but still within the recommended range of 3–5% in terms of mean standardized bias as given by Caliendo and Kopeinig (2008).



culties for older workers in finding dependent employment. The estimated effects for individuals residing in areas with relatively high GDP per capita are slightly larger and may be due to better business opportunities in these regions. Gender differences in estimated effects are small, with the long-term effects being marginally larger for men.

## 6 Sensitivity analysis

In this section, we extensively test the sensitivity of our main results with respect to the implementation of PSM and deviations from the CIA, both within the standard PSM framework and by using an instrumental variable strategy.

### 6.1 Sensitivity with respect to implementation

In order to test whether our baseline results are driven by peculiarities of our chosen matching approach, we check the robustness with respect to more technical details such as the link function  $F(\cdot)$  used to estimate the propensity score, the imposition of common support and the matching or weighting algorithm. These choices have been shown to significantly affect the finite sample performance of estimators (e.g., see Huber et al. 2015; Lechner and Strittmatter 2017). Our findings of this analysis are shown in Panels B to D of Table 4 and can be summarized by stating that none of these discretionary choices regarding the implementation of PSM have any economically significant effect on point estimates. The choice of the link function has very little effect on our estimates, even though the robit regression—which makes use of the  $t$ -distribution with optimally chosen degrees of freedom  $\nu$ —yields quite different predicted propensity scores, as can be seen in Figure A.1 (see Liu 2005, for details on robit regression). The imposition of common support through defining a minimum density ( $\hat{f}(\hat{p}) \geq c$ ) of comparison individuals as done by Heckman et al. (1997) or restricting the analysis to an optimal interval  $[0, \alpha^*]$  as proposed by Crump et al. (2009) yields estimates very close to our baseline estimates.<sup>12</sup> Furthermore, different choices of matching or weighting algorithms also do not play a crucial role for our results. For comparison, we tried pair matching with replacement, radius matching with bias adjustment based on Lechner et al. (2011) and inverse probability weighting with weights re-scaled to unity.<sup>13</sup> The latter two were chosen as they have been found to perform well in Monte Carlo simulations on finite sample properties by Busso et al. (2014) and Huber et al. (2013). Point estimates are very similar across these estimators. While the results are robust to alterations in the design phase of our study, the conclusions drawn depend on the applicability of the CIA, which we aim to assess in the next subsections.

<sup>12</sup> The interval derived by Crump et al. (2009) is optimal in the sense that it minimizes the asymptotic variance of matching estimators. Choosing  $\alpha$  involves a trade-off: Larger  $\alpha$  reduces imbalance and extrapolation leading to lower variance, while discarding information increases variance. As software implementation, we use their accompanying *optselect* package to obtain  $\alpha$ .

<sup>13</sup> The radius matching with bias adjustment is implemented using the *radiusmatch* package of Huber et al. (2015).

## 6.2 Robustness to the inclusion of non-cognitive skills

The recent literature has found that measures of personality traits and non-cognitive skills significantly correlate with labor market outcomes and that non-cognitive skills are about as important in determining wages as cognitive abilities (Heckman et al. 2006). Furthermore, Caliendo et al. (2014) show that personality traits are associated with the decision to become and remain self-employed. Thus, these types of usually unobserved variables are potentially important but omitted confounders that help us to proxy for entrepreneurial affinity. Available through the survey data, we include measures of the individuals' characteristics like the Big Five personality traits, locus of control, risk attitudes, impulsiveness, patience and general self-efficacy in the estimation of the propensity score. Doing so increases the pseudo- $R^2$  of the probit estimation markedly from about 20% to 32%. Thus, differences in personality and non-cognitive skills explain a relatively large part of selection into treatment. For details on estimated probit coefficients, see Table A.2. The resulting propensity score distribution can be seen in Figure A.2. If incorporating these variables into the propensity score estimation significantly changes the resulting treatment effects estimates, this would hint toward a violation of the CIA for our baseline results. However, Panel C in Table 5 shows that the point estimates barely change compared to our baseline estimates and the estimated effects are still highly significant at all conventional levels. These findings are also consistent with those of Caliendo et al. (2016), who analyze the interplay of personality traits and the effects of SUS for a sample of participants of the pre-reform NSUS program with fewer layers of selection at play in much more detail.<sup>14</sup>

## 6.3 Robustness to time-invariant unobserved heterogeneity

The longitudinal nature of our outcome data also allows us to control for time-constant unobserved confounders by means of conditional difference-in-differences (CDID) (see, e.g., Blundell and Costa Dias 2009). The CDID estimator combines the difference-in-differences approach with matching to control for observed characteristics. We choose symmetric differencing around the date of entry, following Chabé-Ferret (2015), who finds that CDID estimators perform the best under this setup.<sup>15</sup> Since the CDID approach requires a weaker form of the conditional independence assumption, this provides a test for the applicability of the original CIA defined in (2). For the CDID estimator to give consistent results, the individual time difference in  $Y^0$  must be independent of treatment when conditioning on the propensity score. Formally, it is required that

$$(Y_{t+k}^0 - Y_{t-k}^0) \perp\!\!\!\perp D \mid P(X), \quad (6)$$

<sup>14</sup> Interestingly, their results suggest a lesser role of personality traits for selection into SUS, which may indeed indicate more severe selection into treatment through caseworkers after the reform.

<sup>15</sup> One additional finding of Chabé-Ferret (2015) is that it is advisable not to condition on pre-treatment characteristics in the matching process when using CDID. However, for our application, this does not make any significant difference.

**Table 5** Sensitivity regarding CIA–PSM

	Outcomes after 20 months		Outcomes after 40 months	
	Self- or regular employment	Net monthly earned income	Self- or regular employment	Net monthly earned income
<i>A. Baseline results</i>	0.280*** (0.000)	764.2*** (0.000)	0.215*** (0.000)	980.2*** (0.000)
<i>B. Including personality traits</i>	0.271*** (0.000)	735.0*** (0.000)	0.201*** (0.000)	915.8*** (0.000)
<i>C. Conditional DiD</i>	0.287*** (0.000)	666.8*** (0.000)	0.255*** (0.000)	914.7*** (0.000)
<i>D. Critical <math>\Gamma^*</math>s for Rosenbaum bounds</i>	12.63	2.73	6.19	2.49

The table shows ATT estimates based on PSM and results from bounding analyses to assess the sensitivity of the estimates regarding unobserved confounders. *p* values for ATT estimates are shown in parentheses underneath the point estimate, obtained by bootstrapping the *t*-statistic using 999 replications (MacKinnon 2006; Huber et al. 2015). \*\*\*/\*\*/\* denote significance at the 1/5/10 % level

Panel A shows our baseline results. Panel B includes usually unobserved personality traits into the score specification. Panel C tests the sensitivity of our results with respect to time-constant unobserved confounders by means of conditional DiD. Finally, Panel D gives the factor  $\Gamma$  by which the odds of receiving treatment between observationally identical individuals must differ to overturn our inference by using Rosenbaum (2002) bounds

where *k* is the number of months before or after (hypothetical) entry *t* into the program. Our results are shown in Panel C of Table 5. As becomes readily apparent, controlling for unobserved time-invariant heterogeneity via conditional differences barely affects our estimates. Thus, the results also do not seem to be very sensitive with respect to this kind of deviation from the CIA.

### 6.4 Assessing sensitivity using bounding analysis

In this part, we follow Rosenbaum (2002) and test the sensitivity of our inference with respect to the degree of departure from the CIA by using a bounding approach. Let  $\Gamma$  denote the ratio of the odds of receiving treatment for two observationally identical individuals *i* and *j*, but different unobserved characteristics *U*, then

$$\frac{1}{\Gamma} \leq \frac{P_i/(1 - P_i)}{P_j/(1 - P_j)} \leq \Gamma, \tag{7}$$

where  $P_i = F(x'_i\beta + \gamma u_i)$  and  $\Gamma = e^\gamma$  for the case of a logistic  $F(\cdot)$ . The bounding exercise essentially varies  $\gamma$  and thus  $\Gamma$  and tests whether the estimated effect remains significant at that level of “hidden bias.” In our application, we assume that we have over-estimated the true effect and gradually increase  $\Gamma$  until we obtain the value

for which our estimates turn insignificant.<sup>16</sup> Thus, if this critical value is large, our estimates are relatively robust with respect to deviations from the CIA. Panel D of Table 5 gives the critical  $\Gamma^*$ s for our four outcome variables of interest. Generally, inference with respect to employment prospect is more robust than earnings. The smallest  $\Gamma^*$  that we obtain is for net monthly earnings after 40 months, which is equal to roughly 2.5, meaning that an unmeasured covariate would need to increase the odds of receiving treatment by the factor of 2.5 compared to someone without this characteristic to turn our conclusions invalid. Hence, our results indicating positive long-term effects on employment and income are very robust with respect to general unobserved confounders.

## 6.5 Estimates using an instrumental variable approach

Should the CIA indeed fail in our application—despite the evidence presented so far—one can still estimate average treatment effects under the condition that we find exogenous variation in the treatment probability. In this section, we aim to estimate treatment effects using an instrumental variable strategy based on regional variation in the likelihood of receiving treatment, using both the standard two-stage least squares (2SLS) estimator and the semi-parametric approach by Frölich (2007). For a dummy instrument  $Z$ , both estimators can be displayed as

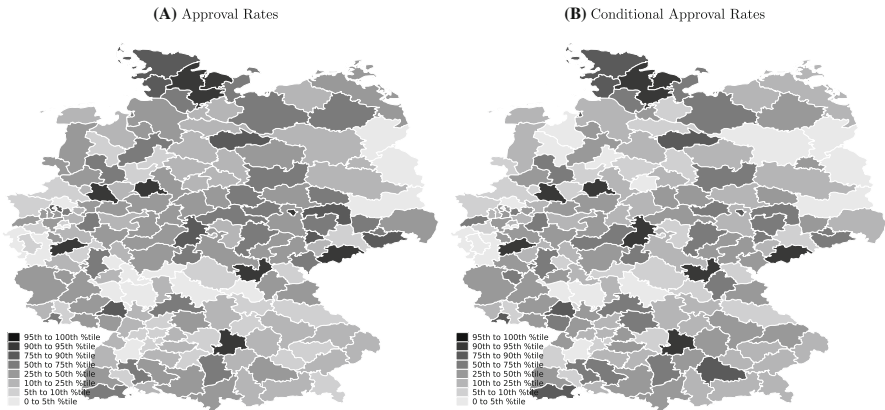
$$\tau_{IV}(x) = \frac{E[Y | X = x, Z = 1] - E[Y | X = x, Z = 0]}{E[D | X = x, Z = 1] - E[D | X = x, Z = 0]}, \quad (8)$$

where 2SLS conditions on  $X$  through linear regression and the IV-matching approach by conditioning on the scalar propensity score  $\tilde{P}(Z = 1 | X) = \tilde{P}(X)$ . The latter estimator is used to test the sensitivity of the IV estimates with respect to the inherent linearity assumption.

For the construction of an instrument, we make use of the regional discretionary power of LEAs after the reform to allocate their allotted funds with respect to different ALMPs. If—conditional on local labor market conditions—an LEA makes stronger use of a certain ALMP, this can be regarded as having a stronger preference for this type of program. Our proxy for the regional preference for SUS is the ratio of approved applications to the total number of applicants in the same region during our sampling time frame, albeit in months other than the month of entry.<sup>17</sup> We call this the leave-one-month-out approval rate, or approval rate for short. Dropping the individual's month of (hypothetical) entry should purge the instrument from a direct relationship with the individual's characteristics. Figure 5 shows the spatial distribution of approval rates

<sup>16</sup> For binary outcomes, we use the *mhbounds* package by Becker and Caliendo (2007), and for continuous outcomes, *rbounds* is employed as described by DiPrete and Gangl (2004).

<sup>17</sup> There are several reasons why we are constrained to contemporaneous data for the instrument. First, data from the previous year correspond to the pre-reform program and thus measure the preference for a nonexistent program. Second, data from the month of January 2012 (the first month after the reform took place) cannot be used as the number of approved applications is contaminated by applicants from before the reform. Third, data after our sampling time frame cannot be used as there was a reform of LEA districts, which led to the disappearance of 22 LEAs. The data on applications for the program and actual entries are obtained from administrative data from the FEA.



**Fig. 5** Regional approval rates. *Note* This figure shows the geographic distribution of the regional of approval rates across the 178 Local Employment Agencies. Panel A depicts the regional distribution of the raw data and Panel B shows the distribution of the conditional approval rates, i.e.,  $E(Z | \text{regional characteristics})$ . Regional characteristics include dummies for northern, eastern and southern Germany and categorical dummy variables for the local unemployment rate, the ratio of vacancies to unemployed, the regional GDP per capita, regional self-employment rates and the start-up rate out of unemployment. The conditional approval rate is interpreted as a measure of the LEA’s preference for start-up subsidies as the ALMP of choice

across the 178 LEAs in Germany, both unconditional and conditional on local labor market conditions.

A number of identifying assumptions need to be fulfilled for our IV approach to give consistent estimates. However, even if the assumptions are true, the IV estimates in general only yield a local average treatment effect for the part of the population that changes treatment status due to the instrument, called compliers (see Imbens and Angrist 1994).

**IV Identifying Assumptions** First, the instrument needs to be *relevant*, i.e., the instrument must have an impact on the likelihood of receiving treatment. It can be assumed that relevance is fulfilled if the instrument has an influence in the first stage, i.e., when the denominator of (8) is significantly different from zero. Second, conditional on  $X$  (or  $\tilde{P}(X)$ ), the instrument  $Z$  must satisfy independence with respect to  $D$  and  $Y$ . Put formally, it is required that

$$Y(z), D(z) \perp\!\!\!\perp Z \mid X, \tag{9}$$

where  $Y(z)$  and  $D(z)$  denote the potential outcomes and treatment status, both dependent on the value of the instrument  $Z$ . This implies that—conditional on  $X$ —the instrument is as good as randomly assigned and it does not have an effect on  $Y$  that does not go through  $D$  (*exclusion restriction*). Third, there must be *no defiers*, implying that  $D(z_1) \geq D(z_0)$  for values of the instrument  $z_1 > z_0$ .

**Plausibility of Assumptions** Apart from the relevance condition, the other identifying assumption cannot be directly tested empirically and needs to be discussed. Assumption (9) is also called the exogeneity assumption in linear regression, and it is often put in terms of correlations: Once we condition on  $X$ ,  $Z$  must not be correlated with relevant omitted factors  $U$  (e.g., entrepreneurial affinity). The credibility of this assumption thus depends on the richness of controls. We largely employ the same specification as described in Sect. 4.1.<sup>18</sup> Since the regional controls hold particular importance in this case, we control for the geographic location of the individual and local labor market conditions. The former are implemented using dummy variables for northern, eastern and southern Germany, while the latter include measures of the unemployment rate, the vacancy-to-unemployed ratio, GDP per capita and—probably most importantly—the start-up rate out of unemployment and the overall self-employment rate using flexible categorical dummies to soak up as much variation due to different local labor market conditions and the local tendency to become self-employed within geographic regions as possible. However, assumption (9) would still fail if LEAs with large *conditional approval rates* tended to select individuals with lower entrepreneurial affinity into treatment. Of course, we cannot test for this. However, we can test for a conditional correlation of the instrument with the previously mentioned personality traits, which are very predictive of receiving treatment. For this purpose, we run an auxiliary regression of the instrument on personal and regional controls as well as our measures of personality traits.<sup>19</sup> The null hypothesis of zero conditional correlation between personality traits and the instrument cannot be rejected at any traditional significance level, supporting the case for the validity of the independence assumption. The exclusion restriction also seems quite natural to us, as we would only expect an effect of regional participation in SUS on individuals' employment prospects and earnings if there are spillover effects of treatment, in which case the local stock of participants would be a relevant regressor in the outcome equation, but not regional approval rates. Furthermore, the no-defer assumption seems plausible since higher approval rates should weakly increase treatment receipt for everyone.

**Estimates** Table 6 provides the results from our IV estimation. Panel A gives the results using 2SLS in combination with a continuous instrument. Panels B and C show results for a dummy instrument, coded as one if the person lives in an LEA with a leave-one-month-out approval rate larger than the median.<sup>20</sup> While panel B still makes use of 2SLS, Panel C uses the IV-matching estimator by Frölich (2007). The latter can be regarded as the ratio of two matching estimators, thus avoiding functional form restrictions regarding the impact of the instrument on treatment probability and outcomes.

First-stage estimates suggest a significant impact of the instrument on treatment receipt. This is supported by the  $z$ -statistics as they are well above the weak instrument threshold of  $\sqrt{10} \approx 3.17$  as given by Staiger and Stock (1997). Focusing on the

<sup>18</sup> The only difference is that we drop interaction terms as these were only included to further improve balance in  $X$  across treatment groups  $D$ . This choice does not affect our IV results in any significant manner. Results with the interaction terms included can be obtained from the authors on request.

<sup>19</sup> Coefficients on personality traits and test results from the auxiliary regression are shown in Table A.3.

<sup>20</sup> The median corresponds to roughly a 50% leave-one-month-out approval rate.

**Table 6** Sensitivity regarding CIA–IV approach

	First stage		Second stage		After 40 months	
	At entry		After 20 months		Self- or regular employment	
	SUS receipt	Net monthly earned income	Self- or regular employment	Net monthly earned income	Self- or regular employment	Net monthly earned income
<i>A. 2SLS (continuous)</i>						
	0.003*** (5.937)	792.7* (1.956)	0.413*** (3.114)	1002.6* (1.905)	0.291** (2.213)	1002.6* (1.905)
		[126.0; 1,459.3]	[-195; 631]	[134.5; 1870.6]	[.075; 507]	[134.5; 1870.6]
<i>B. 2SLS (dummy)</i>						
	0.098*** (4.705)	880.0* (1.709)	0.342** (2.094)	730.5 (1.069)	0.290* (1.764)	730.5 (1.069)
		[33.0; 1727.0]	[-073; 611]	[-393.6; 1854.6]	[.020; 560]	[-393.6; 1854.6]
<i>C. IV-matching</i>						
	0.086*** (3.728)	849.5 (1.162)	0.326 (1.476)	861.4 (0.885)	0.338 (1.466)	861.4 (0.885)
		[-353.1; 2052.1]	[-.037; .689]	[-739.7; 2462.5]	[-.041; .717]	[-739.7; 2462.5]

The table shows first-stage coefficients as well as LATE estimates using the IV approach. z-statistics are shown in parentheses underneath the point estimate. 2SLS standard errors are heteroscedasticity robust. Inference for IV matching is based on standard errors estimated by bootstrap using 999 replications (Mackinnon 2006). \*\*\*/\*\*/\* denote significance at the 1/5/10 % level. 90% confidence bands are shown in brackets. The first stage passes the weak instrument test by Staiger and Stock (1997) if the respective z-statistic is larger than  $\sqrt{10} \approx 3.17$

Panel A shows the baseline IV estimates using 2SLS and the continuous leave-one-month-out (lomo) approval rate. Panel B also makes use of 2SLS but employs a dummy instrument that equals one if the lomo approval rate is larger than the median of the distribution. Panel C uses the same dummy instrument but employs the semi-parametric IV-matching estimator due to Frölich (2007)

2SLS estimates based on the continuous instrument first, we find that a one percentage point increase in the approval rate increases the likelihood of receiving treatment by 0.3 percentage points. Similar to our baseline matching results, second-stage estimates suggest large and positive effects on both employment and earnings. The long-term effect on the probability of being in self- or regular employment is estimated to be about 29 percentage points and thus actually larger than our matching estimates. Effects on earnings are comparable to the matching results, albeit only being significant at the 10% level. Turning to estimates based on the dummy version of the instrument, we find a similar pattern, although the effects are somewhat smaller and less significant when using 2SLS. The first-stage coefficient gives us a direct measure of the size of the complier population, which is estimated to be around 10% when using 2SLS and 8.6% using the IV-matching estimator. Comparing second-stage results, we find very similar results to the 2SLS estimates, albeit they are insignificant due to the higher variability. Hence, the linearity assumption of the 2SLS approach does not seem too restrictive in our application. Overall, our IV approach also suggests positive effects on employment probabilities and earnings for the subgroup of compliers. Making a comparison with our matching results is difficult as the IV approach identifies a different parameter, but matching estimates are included in IV confidence bands.<sup>21</sup>

## 7 Conclusion

In this paper, we provide the first long-term evidence on the causal effects of the new German SUS program for the unemployed after its reform in 2011. The reform significantly altered the institutional setup of the program, leading to ambiguous predictions on the post-reform effectiveness of the subsidy, e.g., due to uncertainty regarding the effects of shortening the first benefit period, during which the bulk of transfers is paid to participants. Our main results based on PSM techniques suggest that effects on employment probabilities and earned income (up to 40 months after entering the program) are positive, economically important and larger compared to the previous program. Thus, there seems to be room for improvement in terms of SUS effectiveness by altering design features of the programs such as the duration of support. The analysis of effect heterogeneity indicates that the program is especially beneficial for older and low-skilled workers.

Since the reform of the program introduced an additional layer of selection by increasing caseworkers' discretionary power, it is necessary to critically assess identification assumptions used to estimate the treatment effects. Hence, we assess the sensitivity of our conclusions with respect to deviations from the CIA within the matching framework and also by using an instrumental variable strategy that exploits regional variation in the likelihood of receiving treatment induced by the reform. Our results of the sensitivity analyses indicate that the matching estimates are very robust to deviations from the CIA and the IV results also point toward large positive effects. Since SUS programs in other countries operate in a similar manner in terms of both

<sup>21</sup> To see this, compare the square brackets in Table 6 with the baseline results from Table 4.



selection and support, our findings of robust positive effects may lend some credibility to other matching estimates in the literature.

While our microeconomic estimation approach provides evidence on individual-level effects of SUS for previously unemployed participants, there may be spillover or general equilibrium effects of such a program that our empirical strategy is unable to identify. On the one hand, SUS programs may have negative spillovers, for example, by displacing other regular businesses due to a competitive advantage of subsidized businesses. On the other hand, SUS may also lead to positive spillovers, e.g., by also leading to job creation for other unemployed jobseekers. Future research should aim to identify these potentially important effects of SUS as this would allow a more thorough analysis of benefits and costs of SUS. Furthermore, it would be interesting to experimentally validate the individual-level results of the observational studies conducted thus far. Given the large number of applications that had to be rejected due to budgetary reasons, this seems a natural way to proceed and would potentially also allow testing different design features (e.g., duration of support). This could help to learn more about the optimal design of SUS for the unemployed.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Human and animal rights** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- Becker S, Caliendo M (2007) Sensitivity analysis for average treatment effects. *Stata J* 7(1):71–83
- Behrenz L, Delander L, Månsson J (2016) Is starting a business a sustainable way out of unemployment? Treatment effects of the Swedish start-up subsidy. *J Labor Res* 37(4):389–411
- Bellmann L, Caliendo M, Tübbicke S (2018) The post-reform effectiveness of the new German start-up subsidy for the unemployed. *LABOUR* 32(3):293–319
- Bernhard S, Grüttner M (2015) Der Gründungszuschuss nach der Reform: Eine qualitative Implementationsstudie zur Umsetzung der Reform in den Agenturen. *Forschungsbericht 4/2015, IAB Nürnberg*
- Blundell R, Costa Dias M (2009) Alternative approaches to evaluation in empirical microeconomics. *J Hum Resour* 44(3):565–640
- Busso M, DiNardo J, McCrary J (2014) New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Rev Econ Stat* 96(5):885–897
- Caliendo M (2016) Start-up subsidies for the unemployed: opportunities and limitations. *IZA World Labor* 200:1–11
- Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *J Econ Surv* 22(1):31–72
- Caliendo M, Künn S (2011) Start-up subsidies for the unemployed: long-term evidence and effect heterogeneity. *J Public Econ* 95(3–4):311–331
- Caliendo M, Künn S (2014) Regional effect heterogeneity of start-up subsidies for the unemployed. *Reg Stud* 48(6):1108–1134
- Caliendo M, Künn S (2015) Getting back into the labor market: the effects of start-up subsidies for unemployed females. *J Popul Econ* 28(4):1005–1043
- Caliendo M, Fossen F, Kritikos A (2014) Personality characteristics and the decisions to become and stay self-employed. *Small Bus Econ* 42(4):787–814
- Caliendo M, Künn S, Weissenberger M (2016) Personality traits and the evaluation of start-up subsidies. *Eur Econ Rev* 86:87–108

- Chabé-Ferret S (2015) Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *J Econ* 185(1):110–123
- Crump R, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1):187–199
- DiPrete T, Gangl M (2004) Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociol Methodol* 34:271–310
- Duhautois R, Redor D, Desiage L (2015) Long term effect of public subsidies on start-up survival and economic performance: an empirical study with French data. *Rev Écon Ind* 149(1):11–41
- Dunn T, Holtz-Eakin D (2000) Financial capital, human capital, and the transition to self-employment: evidence from intergenerational links. *J Labor Econ* 18:282–305
- Frölich M (2007) Nonparametric IV estimation of local average treatment effects with covariates. *J Econ* 139(1):35–75
- Heckman JJ, Robb R (1985) Alternative methods for evaluating the impact of interventions: an overview. *J Econ* 30(1):239–267
- Heckman JJ, Ichimura H, Todd P (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Rev Econ Stud* 64(4):605–654
- Heckman JJ, Stixrud J, Urzua S (2006) The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J Labor Econ* 24:411–482
- Huber M, Lechner M, Wunsch C (2013) The performance of estimators based on the propensity score. *J Econ* 175(1):1–21
- Huber M, Lechner M, Steinmayr A (2015) Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behavior. *Empir Econ* 49(1):1–13
- Imai K, King G, Stuart EA (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc Ser A* 171(2):481–502
- Imbens G, Angrist JD (1994) Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–75
- Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47(1):5–86
- Lechner M (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: *Econometric evaluation of labour market policies*. Physica-Verlag, Heidelberg, pp 43–58
- Lechner M, Strittmatter A (2017) Practical procedures to deal with common support problems in matching estimation. Discussion Papers 10532, Institute for the Study of Labor (IZA), Bonn
- Lechner M, Wunsch C (2013) Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Econ* 21:111–121
- Lechner M, Miquel R, Wunsch C (2011) Long-run effects of public sector sponsored training in West Germany. *J Eur Econ Assoc* 9(4):742–784
- Leuven E, Sianesi B (2003) PSMATCH2: stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Technical report, Statistical Software Components S432001, Boston College Department of Economics, Revised 30 April 2004
- Lindquist M, Sol J, van Praag M, Vladasel T (2016) On the origins of entrepreneurship: evidence from sibling correlations. CEPR Discussion Papers 11562, C.E.P.R. Discussion Papers
- Liu C (2005) Robit regression: a simple robust alternative to logistic and probit regression. Wiley, Hoboken, pp 227–238
- MacKinnon JG (2006) Bootstrap methods in econometrics. *Econ Rec* 82:2–18
- OECD (2015) Public expenditure and participant stocks on LMP. OECD, Stat Database
- O’Leary CJ (1999) Promoting self employment among the unemployed in Hungary and Poland. Working Paper 99-55, W. E. Upjohn Institute for Employment Research
- Perry G (2006) Are business start-up subsidies effective for the unemployed: evaluation of enterprise allowance. Working paper, Auckland University of Technology
- Rodríguez-Planas N, Jacob B (2010) Evaluating active labor market programs in Romania. *Empir Econ* 38(1):65–84
- Rosenbaum PR (2002) *Observational studies*, 2nd edn. Springer, New York
- Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
- Roy AD (1951) Some thoughts on the distribution of earnings. *Oxf Econ Pap* 3(2):135–146

- Rubin D (1974) Estimating causal effects of treatments in randomised and nonrandomised studies. *J Educ Psychol* 66(5):688–701
- Rubin DB (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol* 2(3):169–188
- Sauerbrei W, Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc Ser A (Stat Soc)* 162(1):71–94
- Sianesi B (2004) An evaluation of the Swedish system of active labour market programmes in the 1990s. *Rev Econ Stat* 86(1):133–155
- Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–586
- Tokila A (2009) Start-up grants and self-employment duration. Working paper, School of Business and Economics, University of Jyväskylä
- Wolff J, Nivorozhkin A, Bernhard S (2016) You can go your own way! The long-term effectiveness of a self-employment programme for welfare recipients in Germany. *Int J Soc Welf* 25(2):136–148

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Marco Caliendo<sup>1,2,3,4</sup> · Stefan Tübbicke<sup>1</sup>

Stefan Tübbicke  
stefan.tuebbicke@uni-potsdam.de

- <sup>1</sup> Chair of Empirical Economics, University of Potsdam, August-Bebel-Str. 89, 14482 Potsdam, Germany
- <sup>2</sup> IZA, Bonn, Germany
- <sup>3</sup> DIW, Berlin, Germany
- <sup>4</sup> IAB, Nuremberg, Germany