Discussion Paper No. 16-084

End-of-Year Spending and the Long-Run Employment Effects of Training Programs for the Unemployed

Bernd Fitzenberger, Marina Furdas,
and Christoph Sajons

ZEW
Zentrum für Europäische
Wirtschaftsforschung GmbH
Centre for European
Economic Research

Discussion Paper No. 16-084

End-of-Year Spending and the Long-Run Employment Effects of Training Programs for the Unemployed

Bernd Fitzenberger, Marina Furdas,
and Christoph Sajons

End-of-Year Spending and the Long-Run Employment Effects of Training Programs for the Unemployed

Bernd Fitzenberger*, Marina Furdas**, Christoph Sajons***

December 2016

Abstract: This study re-estimates the employment effects of training programs for the unemployed using exogenous variation in participation caused by budget rules in Germany in the 1980s and early 1990s, resulting in the infamous "end-of-year spending". In addition to estimating complier effects with 2SLS, we implement a flexible control-function approach to obtain the average treatment effect on the treated (ATT). Our findings are: Participants who are only selected for budgetary reasons do not benefit from training programs. However, the ATT estimates suggest modest positive effects in the long run. Longer programs are more effective than shorter and more practice-oriented programs.

Keywords: Training for the unemployed, budgetary conditions, administrative data, Germany

JEL-Classification: J64, J68, H43

---

* Humboldt University Berlin, IFS, CESifo, IZA, ROA, and ZEW.

** Humboldt University Berlin.

*** University of Freiburg and Walter Eucken Institute.

Corresponding Author: Bernd Fitzenberger, Humboldt University Berlin, School of Business and Economics, Spandauer Strasse 1, 10099 Berlin, Germany. E-mail: fitzenbb@hu-berlin.de.

We are very grateful to Stefan Bender, Karsten Bunk, Theresia Denzer-Urschel, Bärbel Höltzen-Schoh, Else Moser, and Georg Uhlenbrock for providing valuable information and to Olga Orlanski for excellent research assistance. This paper benefited from helpful discussions during various seminars, including Café Workshop 2014 Børkop, Labor conference IAB 2015 Nuremberg, CESifo Seminar 2015 Munich, Econometric Conference 2015 Strasbourg, SOLE/EALE Conference 2015 Montreal, EEA Conference 2015 Mannheim, IFS 2015 London, ROA 2016 Maastricht, IFAU Uppsala 2016, and University Odense 2016. In particular, we thank Christian Dustmann, Pierre Koning, Michael Lechner, Artur Lewbel, Olivier Marie, Jeff Smith, Alexandra Spitz-Oener, Joseph Terza, and Jeff Wooldridge for helpful comments. This paper is part of the project "Policy change, effect heterogeneity, and the long-run employment impacts of further training programs" ("Politikänderung, Effektheterogenität und die längerfristigen Beschäftigungswirkungen von Fortbildung und Umschulung", IAB project number: 1213–10–38009). Financial support by the IAB is gratefully acknowledged. This paper replaces the earlier, substantially different working paper "End-of-year Spending and the Long-Run Effects of Training Programs for the Unemployed" which was circulated in 2015.

1 Introduction

The effectiveness of training programs for the unemployed has been a long-debated issue (Heckman et al., 1999; Andersson et al., 2013; Heinrich et al., 2013) and there is renewed interest in active labor market policies in the aftermath of the great recession (Martin, 2015). Proponents of training programs argue that they are a good investment, since enhancing the abilities and skills of unemployed individuals would lead to a quicker reintegration into the labor market, thus resulting in a win-win situation for the unemployed, the government, and the employers. Critics claim that the resources spent on training programs are mostly wasted, however. They argue that the programs themselves do not have any positive impact on later employment, as they keep the unemployed away from the labor market (the so-called lock-in effect). In this view, any observed positive outcome may only reflect a positive selection of training participants.

Most of the literature on the employment or earnings effects of training programs finds a positive impact, but the results are typically based on selection-on-observables identification strategies.[1] It is possible, however, that the assignment to a training program depends upon characteristics of the unemployed which are not observed by the researcher, but influence the individual's employment chances. Therefore, we do not know to which degree these findings reflect the true causal effect. At the same time, studies using exogenous variation in program entry to account for selection on unobservables are quite rare, despite the burgeoning literature on training.[2] This paper attempts to fill this gap using the arguably widespread phenomenon of end-of-year spending as a novel instrument for training participation.

To do so, we exploit the spending incentives created by strict budget rules within the Federal Employment Office (FEO) in West Germany during the 1980s and early 1990s, which caused exogenous variation in program entry. Specifically, local employment offices (LEOs) were subject to two rules: First, their annual budgets were determined primarily based on their spending needs in the previous year. And second, funds allocated to training programs in one calendar year could not be transferred in any way, i.e., neither to finance training programs in the following year nor towards other labor market programs in the same year. Combined, these rules created a strong incentive for local officials to use their

[1] See e.g. the surveys in Heckman et al. (1999),Card et al. (2010),Card et al. (2015), or Martin (2015) and the literature review in section 2 below.

[2] Again, see Card et al. (2010, 2015) and the literature review in section 2 below. The study by Frölich and Lechner (2010) is a notable exception.

whole budget for training programs in each fiscal year. As a consequence, an unemployed had a higher probability to get selected for public-sponsored training in the last months of a calendar year, if her LEO needed to spend down remaining funds, independent of the personal circumstances of the unemployed or her labor market prospects. We therefore use the variation in the financial situation of LEOs in the mid-term review to instrument the individual participation in a training program during the remainder of the year.

With this approach, our paper makes three important contributions to the literature. First, we examine the causal effect of training participation using a novel identification strategy. In particular, we exploit the end-of-year spending behavior of agencies as source of conditional exogenous variation for the treatment probability. Second, we implement our instrumental variable (IV) strategy not only with the conventional 2SLS estimator, but also using a flexible form of the two-step control function approach for random coefficients models described by Wooldridge (2014). Both account for selection with respect to observed and unobserved characteristics, but shed light on the impact of training for different populations of individuals. On the one hand, 2SLS estimates the local average treatment effect (LATE) for the group of compliers, i.e., for those unemployed who participate in training only because unexpected funding is available and needs to be spent. On the other hand, we can use the control function approach to estimate the average treatment effect on the treated (ATT) for all participants. The contrast between these two IV estimates is interesting in its own right, since the impact of training may well be heterogeneous across individuals. If individuals with higher benefits from training are typically assigned first, it could be the case that unexpected additional resources do not produce the employment effects hoped for. Finally, by extending the period of examination after program participation to up to ten years, we provide evidence on the long-run effects of training. These are important, as the costs of assigning an unemployed to a training programs are large and have to be recouped by positive effects on employment for some time after treatment start.

Our main source of data for this study involves administrative records from the German FEO containing the complete employment information up until 2004 for a random sample of 50% of all individuals entering a training programs for the unemployed in Germany between 1980 and 1993, as well as a 3% random sample of all unemployed who did not participate in any such program over the same time period. The data involve spells in unemployment and training as well as spells in employment. We combine these individual level data with information on actual and planned spending from the annual reports of

the FEO. Our rich data enable us to both estimate the size of first-semester deficits or surpluses for each LEO based on actual participation patterns and follow individual employment records up to ten years after treatment start. Furthermore, as we also observe the individuals' labor market participation since 1975, we can additionally control for their employment history prior to the start of unemployment or training.

The first stage estimates show a significant increase in the individual probability to enter a training program after the summer holidays if the respective responsible LEO was thrifty in the first half of the year and had many resources left, i.e., ran a "relative surplus" compared to the other LEOs. The reverse holds for people living in regions whose employment office had a "relative deficit" and thus less funds for the rest of the year than the average. This pattern suggests that spending the entire assigned budget and signaling sustained need for the future was among the goals of the employment offices. Using our instrument, we then provide 2SLS estimates of the employment effects of training for the compliers and estimates of the ATT based on the control function method described by Wooldridge (2014).

Our findings show that there is no positive effect whatsoever of training on the employment chances of the unemployed, when they are assigned to training in order to spent down remaining funds. By contrast, the chances to be employed years after the treatment rise on average by 12 percentage points in the long run for all training participants. This provides evidence that the fit between program and unemployed is significantly worse when the participation originates from unexpected funding for training. Our results provide evidence that training effects differ strongly due to differences in unobservables. Further, we distinguish between different types of training programs. Here, long programs with a strong focus on acquiring new skills fare much better than shorter programs in which the unemployed improve basic skills or work in simulated firms to maintain their general work skills. Thus, the analysis of the merits and costs of training programs should be conducted individually rather than pooling all of them together.

The paper proceeds as follows: Section 2 provides a succint review of the related literature on training. In section 3, we present detailed information on active labor market policy in Germany during the time of our study. Section 4 describes the data. In section 5, we introduce our instrumental variable, the relative surplus in the first half of each fiscal year, and motivate the plausibility of our identification strategy. The basic estimation approach, the construction of our evaluation sample and some descriptive statistics are presented in section 6. Section 7 reports and discusses the empirical results obtained from

2SLS. Section 8 then introduces the control function approach and presents the estimates for the ATT. Section 9 involves the estimation of employment effects for different types of training programs. Section 10 concludes. Further details on the data and the econometric approach can be found in the appendix.

2 Related literature

The literature on the effectiveness of training programs carefully attempts to address the possible selection bias associated with training participation. In short, if only those unemployed participate in a training measure who have good chances to find a job afterwards anyway, comparing the later employment status of participants and nonparticipants does not allow to estimate the causal average effect of the program, but simply picks up the difference in the underlying ability to take advantage of the newly gained knowledge, experience, or skill. (Dynamic) matching approaches in a combination with flexible matching techniques address this problem if the (dynamic) sorting is based on observable characteristics, e.g., if the better educated have a higher probability to participate than the low skilled. They can not account for selection based on unobservable characteristics, however, like motivation, ambition, or discipline.

The empirical evidence on the employment and earnings effects of participation in training based on a selection-on-observables identification strategy shows mostly small positive effects.[3] For the object of our study, training programs starting up to the early 1990s in West Germany, evidence based on large administrative data can be found e.g. in Fitzenberger et al. (2008) and Lechner et al. (2011). Fitzenberger et al. (2008) follow the dynamic evaluation approach proposed by Sianesi (2004, 2008). They estimate the long-run employment effects of training programs for the unemployed in a dynamic context conditional on the starting date of the treatment (treatment vs. waiting) and find both a negative lock-in effect after program start and significantly positive employment effects in the medium and long run. Lechner et al. (2011) estimate the effect of participation versus nonparticipation in training, using hypothetical starting dates for the nontreated individuals. The results are similar, however, suggesting negative employment effects in

[3] In a meta-analysis of 97 international studies conducted between 1995 and 2007, Card et al. (2010) report that training programs seem to be ineffective in the short run, but tend to have positive medium- and long-run effects. See also the surveys in Kluve (2010), Card et al. (2015), and Martin (2015) as well as the literature reviews in the recent studies by Biewen et al. (2014), Heinrich et al. (2013), Osikominu (2013), or Richardson and van den Berg (2013).

the short run and positive employment effects in the long run, with retraining exhibiting the largest positive effect on later employment with about 20 percentage points after eight years.

However, it is possible that the assignment to a training program may depend upon characteristics of the unemployed which are not observed by the researcher, but influence the individual's employment chances (selection on unobservables). To our knowledge, there are only few studies which take this possibility into account.[4] For the U.S., Andersson et al. (2013) estimate the effect of training under the Work Investment Act (WIA) contrasting different non-experimental approaches. They find that difference-in-differences estimation and controlling for firm fixed-effects in a wage regression show little effect on the estimates of training effects. Likewise, Caliendo et al. (2014) use survey data on some variables typically not available in administrative data for Germany such as personality traits, attitudes, and job search behavior. The study finds that some of them are significant determinants of participation in training, but including them leaves the effect estimates basically unchanged. In contrast, controlling for detailed labor market histories is important (see also Biewen et al., 2014, on this point). However, it remains an open question as to whether unobservable characteristics different to those considered by Caliendo et al. (2014) could play a role. This could be addressed using the IV method, but IV studies are very rare in the literature on training programs for the unemployed due to the lack of plausible instruments (see Frölich and Lechner, 2010, who use regional variation in program assignment, for a notable exception).[5]

[4]Again, see Card et al. (2010, 2015). For the U.S., Ham and LaLonde (1996) show the importance to account for selection on unobservables when estimating the effects of training on duration outcomes. In Europe, Richardson and van den Berg (2013) find negative consequences of long training programs in Sweden, while Osikominu (2013) observes positive long-run effects for Germany. Both studies are based on the timing-of-events approach of Abbring and van den Berg (2004) in continuous time, which assumes time-invariant random effects independent of the covariates governing the selection process. Conditional on these random effects, program participation at any point in time is random and only affects exits from unemployment in the future. Aakvik et al. (2005) investigate the impact of Norwegian vocational rehabilitation programs on employment using discrete choice models in a latent index framework with unobservables generated by a normal factor structure and find negative training effects after controlling for selection on observables and unobservables. For Germany, Fitzenberger et al. (2010) use Bayesian techniques in a dynamic framework in discrete time to model selection into and out of training and employment based on observed as well as unobserved characteristics. The study finds positive employment effects.

[5]Frölich and Lechner (2010) examine rich administrative data for Switzerland and exploit exogenous differences of participation probabilities within local labor markets as an instrument for training participation. For the analysis, they use a combination of conditional IV and matching methods to estimate the average treatment effect of participation for compliers.

3 Training programs for the unemployed in Germany

The conduct of training programs for the unemployed has a long history in Germany, dating back to the enactment of the Employment Promotion Act (*Arbeitsförderungsgesetz*) in 1969. This legislation introduced a variety of instruments of active labor market policy (ALMP), with public-sponsored training programs (*Fortbildung und Umschulung*) as the most important component at that time.[6] These programs vary strongly with respect to the intended aim of qualification and their duration, ranging from only few weeks for short-term training to a maximum of three years for complete retraining programs. Their overarching goal is the same, however: To provide general or specific occupational skills in order to improve the labor market prospects of unemployed individuals and those at risk of unemployment. To achieve this aim, the FEO provides financial support for participants which may contain both income maintenance payments and the costs of the program, including money for travel, childcare, and accommodation expenditures. The overall budget available for training programs and income maintenance for program participants totaled around 3.4 billion Deutsche Mark in the mid-1980s (close to 1.7 billion Euros), representing about 11.4% of the annual budget of the FEO at the time.[7]

A detailed description of the various types of training programs can be found in Bender et al. (2005) and Fitzenberger et al. (2008). In this paper, we concentrate our analysis on the four most important ones: Short-term training (*Kurzzeitmaßnahmen*), Practice Firms (*Übungsfirmen*), Provision of Specific Professional Skills and Techniques (*Bereitstellung von spezifischen Kenntnissen und beruflichen Fähigkeiten*), and Retraining (*Umschulung*).[8] In the following, we briefly describe each of these programs in turn, sorted by their average planned duration.

Short-Term Training (STT) courses with an intended duration of no more than six weeks focused on hard-to-place and low-skilled individuals. They were intended to inform job seekers about employment options and possibilities for participation in more comprehensive programs, as well as to provide participants with suitable employer contacts. Furthermore, individuals were taught some general labor market relevant skills, including job search assistance, counseling, and communication training. In general, participants

[6] Other major labor market policy instruments are employment creation schemes (*Arbeitsbeschaffungsmaßnahmen*), promotion of vocational training (*Förderung der beruflichen Ausbildung*), occupational rehabilitation (*Berufliche Rehabilitation*), and short-time work (*Kurzarbeit*).

[7] Own calculations based on figures from the reports of the FEO (Bundesanstalt für Arbeit, 1993).

[8] For the classification of programs, we follow the definitions developed by Fitzenberger and Speckesser (2007).

did not have to take an exam at the end of the course and did not obtain any official certificate at the end (Schneider, 1981). Due to tight budgets after the German re-unification, STT programs existed only until the end of 1992, before they gained importance again with another program design starting in 1997 (Fitzenberger et al., 2013).

Practice Firms (PF) involve a simulated firm environment where participants practice everyday work activities, focusing either on technical or commercial tasks. The program usually lasts six months and aims at providing participants with general skills appropriate for a wide range of jobs. Additionally, PFs are used to assess the participants' ability for particular professions. Similar to STT, participants do not receive a certificate, since the program concentrates on exercising existing skills rather than learning new ones.

Specific Professional Skills and Techniques (SPST) programs focus on providing more specific human capital like computer or accounting courses. The goal of SPST is to facilitate the reintegration of unemployed individuals into the labor market by improving their skills and providing signals to potential employees. A completed vocational training degree is usually required to take part in this type of training. The courses focus on classroom training, but may also provide some practical experience. In case of successful completion, participants usually earn a certificate describing the content of the course and the newly acquired knowledge and experience. Due to the wide variety of courses with durations from several months to up to two years, SPST is the most flexible program and represents the largest share among all public-sponsored training programs.

Finally, the longest and most expensive programs organized by the FEO are *Retraining* (RT) courses. The difference between retraining and the programs described above is that participants actually complete a full vocational training. Most participants in RT already hold a different vocational training degree for a specific occupation, but the prospect of finding a job in that profession is small. Furthermore, RT is also an option for individuals without any vocational degree, provided they meet additional eligibility criteria. In general, it combines both theoretical and practical training, with a total duration of up to three years. After successful completion, participants obtain a widely accepted formal certificate, which serves as a signal for new job qualification.

To qualify for these training programs, unemployed individuals have to fulfill certain requirements, e.g., having worked for at least one year prior to getting unemployed or being entitled to unemployment benefits or subsequent unemployment assistance (Bender et al., 2005; Lechner et al., 2011). Furthermore, full-time enrolled unemployed receive income maintenance payments throughout the duration of their training.

4 Data sources

In this study, we use a unique data set combining information from different administrative sources on training program participation in Germany.[9] The two main components are the *Integrated Employment Biographies* (IEB)[10] and the *FuU* data on program participation. The IEB data are based on administrative *daily* spells reported by employers or the FEO. They contain employment register information for all employees subject to social insurance contributions for the years from 1975 to the end of 2004, thus providing exceptionally long panel data on employment and unemployment periods, benefit receipt from the FEO, and a wide range of personal and job-specific characteristics. Our second data source, the FuU data set consists of *monthly* information about participation in public-sponsored training programs between 1980 and 1997, collected by the FEO for controlling and statistical purposes. This allows to look at further details of program participation and enables a more precise and detailed identification of the exact training measure. After merging the two sources, the identification of training programs is based on combining participation information from FuU data with transfer payment information from the IEB data, giving priority to the former.

As starting point for the construction of our sample, we combine information from a 50% sample of participants in training programs from the FuU data with a 50% sample of program participants from the IEB data. In addition, we use a 3% sample from the IEB data without any program participation as control group, together with individuals who only entered a program at a later point in time.[11] The calculation of average outcomes based on the raw data for aggregates at the local level uses all treated and controls applying the aforementioned sampling weights. Because the 3% control sample is still very large regarding the absolute number of individuals compared to the treatment sample, the relative weight of a treated individual is five times the weight of an individual in the control sample for the estimation of treatment effects. In each data source, we only

[9]The data was generated as part of the project "Policy Change, Effect Heterogeneity, and the Long-Run Employment Effects of Further Training" (IAB project Nr. 1213-10-38009). For the preparation, we used the well documented experience of Stefan Bender, Annette Bergemann, Bernd Fitzenberger, Michael Lechner, Ruth Miquel, Stefan Speckesser, and Conny Wunsch. The main advantage of the new data is its large sample size. While Bender et al. (2005) merely used a 1% sample, the current study is based on a 50% samples for training participants.

[10]In contrast to the standard version of the IEB, our data only contain information from BeH (*Beschäftigten-Historik*) and LeH (*Leistungsempfänger-Historik* of the IAB).

[11]We identify program participants in the IEB data on the basis of transfer payment information, i.e., whether they obtained income maintenance payments of the type that indicates the participation in a training program.

consider information starting in 1980 for reasons of data reliability. Further information on the construction of the data set and the weighting procedure is provided in appendix A.1.

The resulting data set is very informative for two reasons: First, it involves unusually long panel information on employment and unemployment spells from 1980 to the end of 2004, giving us the opportunity to examine the effect of training up to ten years after the treatment. Second, its large size means we have enough observations to conduct the analysis on a local level and to additionally look at the effects of individual programs.

5 "Relative surplus" as instrument

To estimate the impact of training participation on subsequent employment, we need to address various potential sources for selection bias. First, it may be the case that those unemployed individuals are more likely to participate in training who are more able, motivated, and ambitious, and would therefore find it easier to get a job anyway. Second, caseworkers at the local employment offices could base their decisions about a) whether to offer a training opportunity to an unemployed, and b) which type of program seems most appropriate, on their personal assessment of the potential benefits of participation for the respective individual. Third, unemployed may be assigned to training because of their particularly bad employment chances. While the first and second source would lead to a positive selection of participants, the opposite would happen if the third source dominated. In all three cases, we may obtain biased estimates of the effect of training participation on employment for any evaluation relying on a selection-on-observables identification strategy, if we do not observe all information available to the caseworker.

This paper attempts to come closer to estimating causal effects by exploiting budget rules for active labor market policies in Germany which create a source of exogenous variation in training probabilities. In this section, we present the institutional background at the time of our analysis that motivates our instrument and describe the construction of our measure of relative unexpected budget surplus.

5.1 Institutional background

Up until 1994, the most important aspects in the organizational structure of the FEO and its budgeting system for the purpose of this study were the following: (a) The FEO was

organized in three levels, with the central office in Nuremberg, nine regional employment offices at the intermediate level, and 142 LEOs at the lowest organizational level. As depicted in figure 1, the regional employment offices largely corresponded to the states in West Germany, with exceptions for small states, on the one hand, and Bavaria, on the other, which is divided into two regions. (b) The total budget was determined and managed largely by the central office of the FEO, especially for all entitlement programs like income maintenance and training programs for the unemployed (Fertig and Schmidt, 2000). Local offices possessed limited discretion in the use of their allocated funds for training programs. (c) The budget for training programs was planned and allocated separately from other programs like job creation schemes and could not be transferred to other purposes. (d) The allocation of funds top-down to the regional and local offices was based primarily on past levels of program participation, but adjusted for anticipated changes in need across the different regions. (e) Unused funds from one fiscal year could not be transferred to the following year.[12]

Together these budget rules substantially affected the management of training programs for the unemployed by the LEOs. Most importantly, as the budget for the next year depended on the degree of utilization in the current year and its comparison with other local offices, each LEO had an incentive to spend its whole budget before the end of the year, because otherwise it ran the risk of losing funds for the following year. The probability for this was even greater when the other offices did exhaust their budget and could thus plausibly claim their needs for the next year. However, local officials could not simply overrun their allocated resources every year to secure a continuous rise of its funding, as that would have triggered enquiries by the FEO and possibly accusations of inefficient management. As a consequence, the best outcome in the self interest of local decision makers seemed to be to use the whole budget in a year in order to guarantee a stable and possibly growing budget for the next years. Therefore, the degree to which the budget was spent at the time of the midterm review most likely influenced the decision to assign an additional unemployed into training later-on. If budgets were almost exhausted, caseworkers may have hesitated more in assigning further participants. If resources were abundant and needed to be deployed, on the other hand, the chance to be selected to a training program increased.

[12]This institutional framework remained stable up until 1994, when new rules granted a modest level of budget autonomy to the local employment offices. In 1998, a global budgeting system was introduced to make the budget system at the FEO even more flexible.

There is a lot of anecdotal and suggestive evidence for end-of-year spending behavior in government agencies and company divisions (examples include Comptroller General, 1980, Douglas and Franklin, 2006, and McPherson, 2007, for the former, and Merchant, 1985, for the latter), but there is hardly any empirical evidence on its effects due to a lack of reliable data (General Accounting Office, 1998). To our knowledge, the only empirical study is Liebman and Mahoney (2013), who investigate IT procurement decisions of various US federal agencies and find evidence for higher spending on lower quality projects in the last week of the fiscal year.

For our analysis, we use the remaining financial leeway after the first half of the year as a measure for the magnitude of possible end-of-year spending behavior of local officials, which then serves as instrument for program participation in this region in the following months. Figure 2 displays the timing of events in the budgeting process of the LEOs to illustrate this approach. The review of the first semester typically took place in July after all the information regarding the first six months was available. The financial situation at this point relative to the other LEOs then influenced the participation decisions in the remaining months of the year after the summer holidays. Due to a rotation system for the summer holidays between the German states, the "end-of-year" period in our examination therefore starts either in August or September depending on the region and year under consideration and ends in November as the costs for entries in December count towards the January budget.[13]

This strategy is supported by the following two arguments, both originating from personal interviews and correspondence with FEO experts and practitioners. First, the non-transferability of funds between different years caused caseworkers to set program starts as early as possible during a calendar year in order to ensure that available funds were spent during that fiscal year. Second, holiday periods needed to be taken into account for the planning, as many training providers did not offer courses during vacations. This is because the programs required a stable group size and an economically viable number of participants. Therefore, the usual point in time for readjustments in program assignments in response to higher or lower spending during the first semester was after the summer holidays, which ended in August or September depending on state and year.

The evidence in figure 3 supports this reasoning. It shows the average distribution of entries into all considered training programs over the year. We can see that there is a

[13]We treat August as the first month of our examination period for a particular region and year if at most half of the workdays in August belonged to the school holidays, and September otherwise.

pronounced seasonal pattern with many entries during the first semester and a large spike after the summer holidays. In contrast, entry is lowest in June, July, and December. This suggests that caseworkers took the summer holidays as well as the end of the fiscal budget year into account, with training providers accommodating this seasonality of program starts. The next section describes the computation of the instrument in detail.

5.2 The derivation of the relative budget surplus

We follow five steps to construct our instrument "relative surplus": (1) We count the number of program entries in each region for every year and month between January 1980 and December 1993 from our data. (2) We predict the planned entries into training at the local level for each year (i.e., the "budget") based on the entry patterns of the previous three years. This means that 1983 is the first year in our analysis, as the information for 1980 to 1982 is used in the construction of the instrument. (3) We adjust the resulting number by the percentage change in the annual federal budget for training measures to incorporate common macroeconomic effects. (4) We compute a measure for the size of the budget that is still disposable at the time of the mid-term review for each LEO l in year τ as follows:

$$(1) \qquad Budget\ leeway_{l\tau} = \frac{Planned\ entries_{l\tau}(1-12) - Actual\ entries_{l\tau}(1-6)}{Eligible\ unemployed\ in\ July_{l\tau}\ [in 1,000]}\ ,$$

where the arguments (1-12) and (1-6) refer to months 1 to 12 and 1 to 6 in year τ, respectively. We subtract the number of actual entries (1-6) during the first six months in year τ from the planned entries (1-12) for the entire year, corresponding to our predicted budget, and normalize by the number of eligible unemployed in that district in July (in 1,000) as a measure of potential training participants. The resulting budget leeway thus proxies the chance to be assigned to training in the second semester of the calendar year. It increases with the magnitude of the remaining budget and it is positive as long as a local office did not assign more people to training measures in the first half of the year than its total budget for that year. (5) We calculate the average budget leeway for all other offices $s \neq l$ for each year τ and subtract it from the result of region l in that year to obtain our instrument "*relative surplus*":

$$(2) \qquad Relative\ surplus_{l\tau} = Budget\ leeway_{l\tau} - \frac{1}{L-1}\sum_{s \neq l} Budget\ leeway_{s\tau}\ .$$

On the one hand, this controls for common macroeconomic or financial shocks, as they would be equally reflected in the leeway of district l and the average of the other districts, thus canceling each other out. On the other hand, it additionally incorporates the behavioral aspect that local officials in districts with abundant remaining resources may feel a much greater need to spend it if the other districts are already running a tight budget than if they had no financial worries as well. Likewise, LEOs with a high rate of program entries in the first half of the year may not cut back spending too strongly if the situation is the same in the other districts as well, giving them an excuse for not imposing a stricter control over their expenditures.

The resulting value of relative surplus can be positive or negative (i.e., a "relative deficit"), depending on the LEO's position relative to the other offices. It thus proxies the financial leeway a LEO possesses at the end of the summer compared to the other districts and serves as our instrument for starting a training program during the fall.

Two issues of this approach need to be discussed in more detail. One is our use of program entries to proxy the budget. To our knowledge, monthly information on actual and planned expenditures for the 142 LEOs is not available for that time period, so we cannot calculate our instrument from financial figures only. On the other hand, the budget rules at the time in fact involved a direct link between program entries and available funds based on head counts (Bach et al., 1993). Since money allocated to training programs could not be transferred to other ALMP programs (and vice versa), there was no way to spend it on other items. This resulted in a close relationship between annual budgets and the number of participants. Furthermore, the planning and allocation of resources was done for training programs as a whole, that is, not distinguishing between the different program types.

The second issue is how to get the exact numbers for the respective quantities. Based on our 50% sample of all program participants, we can compute the number of new participants by district and year with high precision. However, determining the number of planned entries into all training programs is more complex. Here, we estimate it for each of the 142 districts based on separate out-of-sample predictions. That is, for each LEO and year between 1983 and 1993, we use the data on the number of entries during the three preceding years (in logs) and regress entries in one year on its first lag, the monthly entries and local unemployment rates during the second half of the preceeding year (both in logs), a full set of district dummies, and year as a continuous variable. Then, we use the estimated coefficients together with the actual data to predict the annual entries for the

current year. Note that the model used for the prediction of planned entries is estimated separately for each year τ and the model estimates are only based on past data for the years $\tau - 3$ to $\tau - 1$. The predictions for year τ therefore exclusively rely on information available at the end of year $\tau - 1$.

To increase the precision of this prediction and strengthen the link between our forecasting procedure and the ex-ante planned budget, we account for foreseeable changes in the total number of program participants on the national level by including a correction for changes in planned spending on training programs at the federal level, as indicated above in step 3. For this purpose, we calculate the ratio between intended spending in a certain year, say 1988, and actual spending in the previous year, here 1987. A ratio less than one implies a planned reduction of overall money for training programs and likewise in the absolute number of entrants compared to the preceding year. For the analysis, we multiply the raw predictions of the absolute number of entries with this ratio. The resulting variable does a very good job in explaining the variation in the real annual entry rates, with an R^2 of 81.97 in a regression of actual entries on our predicted ones.

5.3 Validity of the instrument

Figure 4 reports the distribution of relative surplus. By construction, it has a mean of basically 0, while the median is negative at -4.65. This means that the distribution of relative surplus is slightly skewed to the right and that the budget leeway for the second semester was a bit smaller in the median LEO than the average of the other districts. In other words, the median district spent a bit more than the average in the first semester, which is plausible if local officials care about showing that they actually need all of their budget. Compared to the overall monthly average of 8.45 entries into training programs per 1,000 eligible unemployed during the whole time period, this "excess spending" over the first half of the year amounts to the typical inflow of about two to three weeks. Additionally, figure 4 shows a large variation in relative surplus across region and year. While the majority of observations lie between -30 and +30 from the national average (the standard deviation is 34.5), some LEOs even experienced relative surpluses or deficits of more than 70 in some years, with extreme values of up to -95 and +167 at the tails of the distribution. As a consequence, some LEOs are very likely to expand training programs after the summer holidays in order to spend their whole budget, whereas others needed to restrict their spending substantially if they wanted to prevent exceeding their funds and

getting reproached for that. In section 7.1, we show that this is in fact the case. Thus, the budget leeway significantly affects the probability of eligible unemployed to be assigned to a training program during the remainder of the year.

Exclusion restriction

Apart from possessing a relevant impact on the treatment probability, it is crucial for the validity of our approach that our instrument does not directly influence the outcome variable, i.e., that relative surplus only affects later employment through its impact on the treatment probability. Put differently, it is necessary that relative surplus contains a high degree of randomness and is not influenced by any omitted variable that may simultaneously impact the respective district's labor market in the long-run. So what could be potential threats to the validity of our instrument and how do we deal with them?

First, and most importantly, the budget surplus may proxy for local labor market conditions. Specifically, a budget surplus may be associated with particularly good local labor market conditions. To account for this possibility, our specification of the employment equation controls for differences in local labor market conditions by including several variables capturing the level and dynamics of local unemployment. These include the average local unemployment rate from the previous year as well as the monthly unemployment rates for each month from January to July in the current year.[14] Additionally, we also control for persistent differences across regions. Second, the LEOs may differ in the quality and style of their management, as well as by their seasonal pattern of demand for training programs, which could explain some variation in the instrument. To take this into account, we also include the relative surplus in the previous year in the employment equation, thus netting out any potential differences between the managements of the districts. Third, developments at the federal level could influence both the LEO's budget situation and the employability of participants. This should not be problematic in our approach, as we compare each single district with the other LEOs in the construction of the instrument and additionally introduce year fixed effects in the regressions. Thus, any macroeconomic shock that affects the districts more or less equally does not harm our identification strategy. The same holds for any unexpected change in the organization of the federal employment office or across-the-board adjustments in spending on training

[14]Unobserved personal characteristics of the unemployed should not play a role here, although they may strongly influence employment. This is due to the construction of relative surplus at a more aggregate level, where there is no obvious link between the instrument and unobserved individual characteristics.

programs which could be related to the national economic situation. And finally, we control for persistent level differences and diverging time trends between regions by including regional fixed effects and interactions between year and region dummies.

After adding the aforementioned controls, the remaining conditional variation in our relative surplus variable reflects unexpected changes in the micro-environment of each LEO. On the demand side, these may include unanticipated shifts in the type of unemployed and their need for training. On the supply side, we may see changes of personnel or absences of caseworkers in the employment office, as well as changes in management and policy style. Furthermore, local training providers may adjust their services and how they match with the interest of the unemployed and local employers. And finally, the relative surplus variable may also be affected by changes in the majority of the other LEOs, which are not under the control of the officials in an individual district. As all of these sources are sporadic and unsystematic, we do not observe extended periods of large relative surpluses or deficits for any employment office in our data once we condition on our set of control variables. This conclusion is supported by the statistical finding that the residuals of our conditional instrument, after controlling for the other covariates in the employment equation, are not autocorrelated over time, i.e., shocks to the relative surplus in one year are already absorbed in the following year.

6 Implementation and descriptive statistics

6.1 Construction of sample and estimation approach

For the implementation of our IV identification strategy, we define our sample to achieve two goals: On the one hand, we want to restrict our analysis on the group of unemployed individuals who are eligible for participation in training programs and who are the main target group. We do not consider programs related to youth unemployment, early retirement, or the reintegration of former inmates, and we focus on individuals living in West Germany who are between 25 and 50 at the beginning of their unemployment spell, have worked for at least three months before losing their job, and receive either unemployment benefits or assistance within the first quarter of unemployment at the latest. Furthermore, we exclude observations that enter unemployment at a certain point in time and remain

without employment for more than 72 calendar months due to issues of data quality.[15]

We need to account for the dynamic sorting of individuals into the treatment with respect to elapsed unemployment duration in order to ensure the comparability with the members of the comparison group of nontreated individuals. To do so, we use a dynamic risk set matching approach following Li et al. (2001) and Sianesi (2004, 2008).[16] To implement it, we estimate flexible panel employment regressions. In contrast to the literature, we choose to implement risk set matching in a regression setting because this allows naturally for IV estimation.

For a random sample of unemployed, who are eligible at elapsed duration $el = 1, ..., 12$ months, we define treatment as the first participation in training programs during the first 12 months of the respective unemployment spell. Risk set matching compares a person entering a training program after a certain elapsed duration of unemployment only to those observations who became unemployed in the same month, who are still searching, and who have not started training yet. The latter comparison group thus includes those who either never participate in a training program or who start it later. To implement the risk set matching approach in regression analysis and construct a sample by potential treatment starts, all individuals are replicated for each month they remain unemployed and are eligible for treatment or just start one. When we estimate the outcome equation for employment by month since treatment start, we account for the fact that the composition of treated and nontreated individuals depends upon the elapsed duration in months. For example, if an individual starts a treatment in month 7 of the unemployment spell and does not find a job before month 8, the individual is used as control observation for treatment starts during months 1 to 6 and as treatment observation for month 7. If an individual enters unemployment more than once during our period of observation, she is part of different unemployment cohorts and thus appears several times in the empirical analysis. Through the alignment by elapsed unemployment duration, our estimation approach estimates the average treatment effect for the treated, who start treatment in month el of their unemployment spell, which would be the ATT under a selection-on-observables assumption. The average effect estimates for participation during the first twelve months of unemployment average across the 12 months of potential treatment starts, which are weighted by the monthly number of entries into treatment. When

[15] In particular, many of these long-term unemployment spells may be caused by gaps in the employment history, which are considered as non-employment in the data.

[16] See also Fredriksson and Johansson (2008) for a related approach and Biewen et al. (2014), Fitzenberger et al. (2013), or Lechner et al. (2011) for applications.

estimating a conventional two-stage-least-squares regression in this setting, we estimate the LATE among the compliers, i.e. those treated whose treatment status is changed by the variation of the instrument.

Using the risk set matching approach, we effectively estimate the effect of training versus waiting, where non-treatment up to a certain elapsed unemployment duration *el* involves the possibility to be treated later during the course of the unemployment spell. Our analysis starts with those observations who entered a training program in August 1983, therefore including individuals who became unemployed between September 1982 and July 1983. We cluster the standard errors in our regression analysis at the district and year level because the instrument only varies between district-year cells.

6.2 Descriptive statistics

Table 1 provides the relevant sample sizes of treatment and control group by length of unemployment at program start. It shows that the absolute number of entries into training programs declines with a longer duration of unemployment, reflecting the fact that the number of individuals in a cohort falls as more and more of them find a job. At the same time, the fraction of people who start a program increases continuously from 4.8% in the first quarter of unemployment up to 6.1% in the fourth quarter.[17] This may reflect that both the unemployed and their caseworkers find such programs more necessary the longer unemployment lasts. Additionally, we also show the number of participants in the different types of training measures to provide a sense of their relative importance. In particular, we can see that the diverse training programs summed up in the category SPST (provision of Specific Professional Skills and Techniques) are the most frequently used, accounting for more than half of all the program participations (51.7%). The remainder is almost evenly split between the long-running retraining (26%) and the two short measures short-term training and practice firms (22.3%).

Table 2 involves a comparison of individuals in the treatment and control group with respect to the main personal characteristics like gender, age, education, and marital status, but also considering very detailed information about previous work and training biographies. This includes previous employment status, wage, the type of employment, the occupation, and the size of the last firm worked in, as well as whether one partic-

[17]Note that these unweighted numbers reflect the situation in our evaluation sample in which treated individuals are over-represented by a factor of 182 to 12, see section 4. The fraction of program starts is much smaller in the population, but the relative changes are correctly represented in our sample.

ipated in a training program within the last year or two years before. Bold numbers indicate statistically significant differences in the averages between the two groups. Thus, it turns out that participants differ significantly from nonparticipants regarding almost all characteristics considered. For instance, participants are 5 percentage points more likely to be in the youngest age group (25-29 years) than nonparticipants, but 6 percentage points less likely to be in the oldest (45-50 years). This is consistent with training being a human capital investment, paying off the longer the younger the age at the time of training. Likewise, we observe smaller fractions of married individuals and foreign citizens among the treated in all programs compared to the control group (44.9% to 47.4% and 7.1% to 11.8%, respectively). This fits to the investment logic, as married individuals are older on average than singles, and foreigners may consider a return migration, thus having a shorter expected payoff period. Apart from that, the participants are also more likely to be female than the nonparticipants (44.9% to 40%) and earned slightly higher wages in previous employment (3.83 to 3.79), in which they were more likely to work as a white-collar worker (41.6% to 31.6%), mostly in the service sector (59.7% to 53.2%).

Looking at the different types of training programs individually, we see some indications that they focus on different groups of unemployed. For instance, the age pattern described above is most pronounced for retraining, the longest and at the same time most expensive program with a duration of around three years. Likewise, we observe a lower share of individuals with at least some college education in short-term training and practice firms, as these are mostly focused on acquiring and maintaining more basic and practical skills.

Overall, these differences both between participants and nonparticipants and across programs reveal a strong selection with respect to observed characteristics. This highlights the challenge faced by an identification strategy based on the selection-on-observables assumption. If participation follows an investment logic as discussed above, it is likely that the returns to training differ systematically not only by observed, but also some unobserved characteristics (see Card, 2001, regarding heterogeneity in returns to education).

7 OLS and LATE employment effects

This section presents OLS coefficients and standard IV estimates based on 2SLS, as well as results from a reduced form model using our budget instrument as main explanatory variable. 2SLS estimates the local average treatment effect (LATE) of training for the

compliers, who are those individuals who would not have taken part in any program in the absence of a spending rush at the end of the budget year. Note that OLS and LATE estimates may differ because of selection on unobservables or because compliers and always-takers react systematically different to the treatment. We explore this issue further in section 8 using the control function method for random-coefficients models described by Wooldridge (2014), which enables us to estimate the average treatment effects for the treated (ATT).

7.1 First stage: Effect of surplus on participation in training

The first step in the analysis is to check whether our instrument has predictive power for the probability to start a training program, i.e., whether it is actually relevant. In 2SLS, we run a first stage OLS regression of the probability to enter one of the four considered programs in the months after the summer holidays on the size of a district's relative budget surplus. As described above, we control for the first lag of the relative surplus, personal characteristics, work biographies, indicators for each region and year, their interactions, and the development in the local unemployment rate in order to focus on the pure "surprise" effect of relative surplus.[18] We estimate pooled monthly regressions conditional on still being unemployed and not having entered a training program before the month of interest.[19]

The resulting standardized coefficients for a one-deviation change in relative surplus are reported in table 3, together with the corresponding clustered standard errors (at the district and year level) and F-test statistics.[20] The estimates show that the probability to participate in the final months of the year reacts to our measure of relative surplus during the first semester. An increase of one standard deviation in the relative surplus leads to a modest, but significant, change of 2.6 percentage points (ppoints, rounded to the first decimal point) in the participation probability in the first specification, in which we control for regional and year fixed effects as well as a large number of personal

[18] Appendix A.3 provides a detailed description of all explanatory variables used in the analysis.

[19] Since the share of the treatment group is very small in the full sample and we estimate the treatment probability by elapsed unemployment duration, 2SLS may lead to noisy and unstable results (see Chiburis et al., 2012). For this reason, we increase the weight of the treatment group relative to the control group by a factor of five, and we do not apply population weights (see Solon and Wooldrige, 2015, for a discussion of weighting in regression analysis), except for reweighting the treated and nontreated in the comparison group used in the risk-set matching approach, see section 4 and appendix A.1.

[20] Estimating the same equation with a non-linear Probit model yields average partial effects of the same sign and order of magnitude.

characteristics of the unemployed. This effect decreases a bit when we sequentially add the first lag of relative surplus, region and year interactions, and controls for the local unemployment rate over the previous months and the past year, but consistently remains positive and significant. In our preferred and most conservative specification (column 5), a marginal change of one standard deviation in relative surplus (roughly, 34.5 in absolute value) leads to an increase in the participation probability of 1.9 ppoints. Relative to the average treatment probability in our weighted sample of 65.8%, this represents an increase of 2.8%.

This result is also robust to changes in the way we compute the instrument. In particular, if we vary the number of years of information used to predict the number of program entries to either two or four, we still obtain coefficients of the same order of magnitude (1.8 and 1.6, respectively) and significance. The findings therefore indicate that officials at the employment offices reacted to the budgetary environment with the aim of spending unused funds by the end of the year (or to meet the budget in the case of deficits). The corresponding F-statistic for the instrument is large and in all cases way above 10, suggesting that relative surplus can be considered as a strong instrument.

7.2 Employment effects: OLS, reduced form, LATE

The estimated employment effects are reported in table 4. For simplicity and to gain precision for the IV estimates, our outcome variable is the employment rate, i.e. the average the month-specific employment dummies, for three time periods: The first year after the start for the immediate short run, the second and third year capturing the medium run, and years 4 to 10 for the long-run effects. For comparison, we report the outcomes of different estimation methods for both specification 4 and 5 of the first stage, where the difference lies in the use of the local unemployment rate in the previous year as additional control. In column 1, we begin by stating the raw descriptive difference in having a job between treatment and risk set comparison group observations. It starts with a large difference of almost 20 ppoints in the first year after entering the training program, which is as expected as participation reduces the job search activities of unemployed individuals (the lock-in effect). This gap is still present in the second and third year, but it shrinks already to only -5.4 ppoints, as most of the shorter programs have ended. In the long-run, however, former participants in a publicly-sponsored training program are on average 2.3 ppoints more likely to be employed than nonparticipants.

These results are almost perfectly mirrored in columns 2 and 3, displaying the coefficients of an OLS regression of employment on a participation dummy and our extensive set of controls.[21] Columns 4 and 5 show the coefficients of reduced form regressions, where we simply replace the binary participation dummy by our continuous instrument. The sizes of the coefficients are thus not comparable to the previous ones any more, as they indicate the change in employment rates for a marginal increase in relative surplus. Nevertheless, the signs already give us an indication of the direction in which the IV estimates should go. Interestingly, we observe a negative sign for the reduced form even in the long-run period, questioning an overall positive effect of training for the compliers, i.e., those unemployed who start training due to the respective LEO's need to spend down resources, but who would not have done so if less funding was available.

The 2SLS results based on two specifications are stated in columns 6 and 7. Column 7 shows our preferred specification which uses further controls for the local labor market conditions. At first glance, the coefficients show worse employment effects compared to OLS. While this is not the case in the first year after treatment start, the estimate amounts to -10 ppoints during in the medium-run and, most importantly, to -3 ppoints the long-run effect (in column 7). This suggests that the effect of taking part in a training program on the compliers is unlikely to be positive. However, the 2SLS estimates are not statistically significant and somewhat imprecise, which is also reflected in the differences between the two specifications in columns 6 and 7.

8 Average treatment effect on the treated

2SLS estimates the average training effect for the compliers, which is identified under fairly weak assumptions, if a suitable strong instrument is available. However, with heterogeneous treatment effects the complier effect may differ strongly from the ATT, which is typically the focus of studies in the training literature building on a selection-on-observables strategy. In our specific application, the complier group involves individuals who only participate in training if there are more funds available than expected earlier. If training is well targeted on those with the highest returns of training and individual returns are heterogeneous for unobservable reasons, we would expect that the ATT exceeds the LATE. Furthermore, the 2SLS estimates in this paper are imprecise and often

[21] Recall that any point estimate discussed here represents the weighted average of the separate estimates by month of program entry during the first year of the unemployment spell.

not significant. Estimating the ATT requires stronger modelling assumptions compared to the LATE. However, under appropriate modelling assumptions, the ATT may be obtained with more statistical precision than the LATE. To derive the ATT, we implement a flexible two-step control function estimator for a random-coefficients model with both a binary endogenous treatment and a binary outcome (see Wooldridge (2014)). As before, we focus on the employment effect of participation in a training program compared to non- or later participation. In the following, we sketch our implementation of the estimator and refer to appendix A.2 for a formal description.[22]

8.1 Estimating the ATT

To implement a random coefficients model with both a binary endogenous treatment and a fractional outcome variable (the employment rate of an individual within the three time periods), we adopt a flexible control function approach for non-linear models with discrete explanatory endogenous variables as described in Wooldridge (2014, section 6). Wooldridge (2014) shows how the model can be used to obtain the average treatment effect (ATE). We extend the approach to estimate the average effect of treatment on the treated (ATT). The model accounts for a random coefficient of the training dummy, which reflects the heterogeneity of the treatment effect unrelated to exogenous covariates.

Define $d = 1$ to indicate participation in training and $d = 0$ otherwise. According to the risk set matching approach, we align individuals who start treatment in a certain month of their unemployment spell with all individuals who have not started treatment by that month, as measured by elapsed unemployment duration.[23]

In the first stage, we estimate pooled probit regressions for the treatment dummy d for entries into training programs during the first twelve months of unemployment. As covariates z, we use the regressors of the employment equation z_{i1} and our budget instrument z_{i2}. Then, we take the resulting coefficients to calculate the generalized residuals $\widehat{gr}$ as the treatment-status specific inverse Mills' ratios.

For the second stage, we estimate fractional probits for employment, where the dependent variables are employment rates during year 1, years 2 to 3, and years 4 to 10,

[22]For the results reported above, we use regression models which include an indicator variable for treatment, but no interaction between the treatment dummy and observable covariates to account for possible heterogeneity of the treatment effect. We re-estimated the OLS regressions with such interactions and calculated the ATT explicitly as an average partial effect for these regressions. The results of this exercise largely coincide with our former coefficient for the treatment dummy without interactions. The detailed outcomes are available upon request.

[23]Note that the discussion here omits time indices for the month of treatment start or calendar time.

calculated as the time averages of the monthly employment dummies. We account for selection into treatment status by adding $\widehat{gr}$ and interactions of $\widehat{gr}$ with z_{i1} as control functions. Furthermore, our model allows for a random coefficient of the training dummy in the employment regression, which accounts for possible heterogeneous treatment effects. Finally, we routinely test for significance of interactions involving $\widehat{gr}$ as part of the specification search.

We estimate the following fractional probit for employment with control functions based on the following specification of the expected employment rate in period t:

$$(3) \qquad \widehat{E}(y_{it} \mid d_i, z_i) = \Phi\left(\hat{\delta}_{0t} + \hat{\delta}_{1t} d_i + z_1 \hat{b}_0 + z_1 \hat{\delta}_1 d_i + \hat{\omega}_0 \widehat{gr}_i + \hat{\omega}_1 \widehat{gr}_i d_i + z_{i1} \widehat{gr}_i \hat{\psi}\right),$$

where t corresponds to year 1, years 2 to 3, or years 4 to 10 after treatment start, $\Phi(.)$ is the standard normal distribution function, $\hat{\delta}_{0t}$, $\hat{\delta}_{1t}$ are time specific effects, and $\hat{b}_0$, $\hat{\delta}_1$, $\hat{\omega}_j$ $(j = 0, 1)$, $\hat{\psi}$ further coefficient estimates. Appendix A.2 provides a formal discussion of the specification in equation (3).

Based on the results of equation (3), we then derive the ATT by integrating out the distribution of the covariates and the control function terms among the treated $d_i = 1$ as:

$$(4) \qquad \widehat{\tau}_{ATT,t} = \frac{1}{N_1} \sum_{d_i=1} \Big\{ \Phi\left(\hat{\delta}_{0t} + \hat{\delta}_{1t} + z_1 \hat{b}_0 + z_1 \hat{\delta}_1 + \hat{\omega}_0 \widehat{gr}_i + \hat{\omega}_1 \widehat{gr}_i + z_{i1} \widehat{gr}_i \hat{\psi}\right)$$

$$- \Phi\left(\hat{\delta}_{0t} + z_1 \hat{b}_0 + \hat{\omega}_0 \widehat{gr}_i + z_{i1} \widehat{gr}_i \hat{\psi}\right) \Big\} .$$

The first term in the difference denotes the employment probability for the treated, the second for the nontreated. To estimate the ATT, the two terms have to include the individual-specific control function corresponding to the treatment state considered.

Inference when estimating the fractional probit with control functions and the ATT is based on a cluster version of the weighted bootstrap (see Barbe and Bertail, 1995, and Fitzenberger and Muehler, 2015). We cluster standard errors at the district and year level (see appendix A.2 for further details).

8.2 Results

Specification tests

We first use the data to determine appropriate specification of our control function. For this purpose, we start with a flexible model that includes not only the generalized residuals

obtained from the first stage, but also interactions between the generalized residual (gr), the treatment dummy (d), and selected important covariates (X). We then test separately whether the components of this general setup are (jointly) significant to obtain a model that best explains the variation in employment probabilities. The results of these tests are summarized in three different specifications in table 5, all of which use the set of control variables of our benchmark model for the 2SLS results above.

In the first panel, CF 1, we start by including the generalized residuals and an interaction with the treatment dummy (with coefficients ω_0 and ω_1, respectively). Here, we see that the two components are not individually significant at the same time in either period under consideration. The results of χ^2-tests, however, show that they are jointly important both for the first year and the long run, indicating a high degree of endogeneity in the participation in training (the corresponding p-value is 0.000 in both cases). In CF 2, we check whether we can reduce the importance of the selection terms by using the information contained in the observable characteristics more flexibly. To this end, we additionally include interaction terms between individual characteristics and the treatment dummy. It turns out, however, that this is only the case in the first year after program start, but starting with the second year, we find strong and significant indications for selection effects (with p-values of 0.002 and 0.000 for the second and third period, respectively).

Finally, CF 3 adds a set of interaction terms between generalized residuals and individual characteristics (with the vector of coefficients ψ) to account for a possible heterogeneity of selection effects. Here, the results of the two χ^2-tests show that each set of variables is jointly significant. This means that we not only have strong endogeneity in program participation, but also that this selection varies by characteristics of the individual. To incorporate this finding in our analysis, we therefore use CF 3 as our preferred specification for the estimation of our second stage.

Second stage results

Tables 6 and 7 report the results of the full-fledged control function approach, differing only in whether we control for the local unemployment rate in the previous year or not. In both cases, we state the estimates from our linear OLS and 2SLS models from before as reference points.[24] Then, we display the result of a probit model in order to show whether

[24]Note that OLS ATT estimates using interactions between the treatment dummy and exogenous controls $z_{1i}d_i$ as in equation 3, fractional probit estimates, and semi-parametric estimates for the ATT based on inverse probability weighting basically coincide with the effect estimates from the OLS regressions estimating a uniform treatment effect, which are reported here. These further results are available upon request.

including possible non-linearities already leads to different results. Finally, we present the ATTs of the control function approach obtained from each of the three specifications discussed above, with CF3 as the preferred one.

Starting with table 6, three aspects seem to be particularly noteworthy: First, there is no difference between the various methods with respect to the pattern of results they produce. That is, all of them exhibit significant lock-in effects in the first year, slightly better outcomes in the second, and even better results in the long run. Second, the estimates obtained by selection-on-observables approaches (OLS and probit) are almost identical, but differ strongly in the long run from those of the two IV approaches, although we control for a large number of relevant explanatory variables. This shows that non-linearity is not much of an issue here and provides another illustration for the presence of strong selection into program participation and the great challenge of taking it into account. Third, the comparison between 2SLS and CF estimates indicates two important differences. On the one hand, the flexible specification of the selection correction yields more precise estimates than the traditional 2SLS specification. In table 6, the resulting standard errors are smaller in the control function approach by around 45% on average over the different examination periods. On the other hand, we can see that the point estimates deviate quite strongly between the two approaches. These differences are not significant in the first and second period, but in the long run, we have a negative, but insignificant 2SLS result of -8.3 ppoints in contrast to a significant positive effect of 10.1 ppoints in our preferred CF specification.

The last point highlights the difference in meaning and interpretation of the coefficients in the two approaches again. While the CF method estimates the average effect on all the treated individuals, standard 2SLS only measures the impact on the subpopulation of individuals who only took part because the respective LEO needed to spend their remaining funds, independent of whether that makes any sense for them. Thus, we would expect the LATE estimates to be smaller or even negative compared to the ATT ones.

In table 7, we check the robustness of these results by controlling in addition for the local unemployment rate in the previous year. This accounts for possible local unemployment dynamics which could influence both the current relative surplus and later employment probabilities. Here, we see exactly the same pattern as in table 6, with the only difference that the estimates of 2SLS and CF get both a bit more positive. Thus, while the coefficient in the 2SLS model is still negative, it is now only at -3 ppoints in the long run. Similarly, the coefficient of training increases slightly for CF3 to +12.4 ppoints.

Summing up, we see that unemployed individuals modestly benefit from participating in a public-sponsored training program (with about 1 out of 10 participants getting a job because of the program), but this effect does not materialize if they are mainly selected for budgetary reasons.

The results reported in tables 6 and 7 can also be used to discuss the empirical question of whether participants in training programs are negatively or positively selected on unobservable characteristics. That is, do caseworkers tend to send mostly lower motivated, less organized and disciplined unemployed or is it the other way round? To answer this question, we look at the differences between the OLS and the CF estimates, as the former do not control for unobservable characteristics while the latter take their influence into account.[25] Concentrating on the long-run effects, we observe that the CF coefficients are higher than those obtained from OLS. This means that participants tend to be negatively selected with respect to variables which are typically unreported, but affect an individual's chances to find a job.

9 Heterogeneous effects across training programs

As described above, the programs considered in the analysis so far differ vastly in several dimensions, most importantly with respect to duration and focus on acquiring new skills or maintaining old ones. Therefore, estimating a simple average effect for all of them together may hide a large degree of heterogeneity between them. In this section, we want to see whether this is the case, i.e., we want to evaluate their individual impact separately.

The econometric challenge in this context is that we have only one variable, relative surplus, to instrument four potentially endogenous participation decisions at the same time. Therefore, we estimate the effect of participation in one type of training in a certain month of unemployment against the alternative of not participating in training at all in that month, thereby excluding those individuals who participate in other training programs.

Thus, our approach differs from Heckman et al. (2008) who suggest IV estimation in settings with multiple unordered treatments, where the estimation of pairwise effects of one treatment versus another requires the availability of as many instruments as there

[25] We do not consider the 2SLS results in this respect, as we want to make a general point and not one about the limited subset of compliers with our instrument.

are treatments.[26] As we do not focus on evaluating pairwise effects of different training programs, we abstain from estimating an employment equation which models the effect of all training programs simultaneously, i.e., includes multiple endogenous treatment variables in one outcome equation. We have only one instrument which potentially affects participation in all training programs and therefore the respective non-treatment group may change its selection in response to changes of the instrument. Therefore, we focus on the binary comparison of participation in one type of training versus nonparticipation in any training program.

This approach is useful for two reasons. On the one hand, we estimate the first-stage selection model in the CF approach for each training program separately based on those participating in the respective program and those not receiving any treatment. This accounts for the selection of the treated relative to the control group regarding the non-treatment outcome, which is what we need for the estimation of the ATT. On the other hand, the treatment probability for all training programs in a certain month is very small (recall that the treated are overrepresented in our evaluation sample), which means that the composition of the non-treatment group is hardly affected by the effect of our budget instrument on the participation in the alternative training programs. Thus, the key issue for the estimation of the ATT is to control the selection bias among the treated, which is what our approach focuses upon. Note that our instrument would change the selection of both the treated and the controls if we wanted to estimate the pairwise effect between two different programs. Thus, in light of Heckman et al. (2008), it would be difficult to justify and we consequently refrain from doing so.

Table 8 reports the normalized individual first stage results obtained from a linear probability model. The corresponding average partial effects from a probit are not included, but are very similar and can be obtained upon request. We find that an increase of one standard deviation in relative surplus raises the probability to take part in any of the four programs significantly, with a range between 1.1 ppoints for retraining and 2.7 ppoints for practice firms. Relative to the average participation probability in the evaluation sample, this corresponds to an increase of between 17.4% for practice firms and 3.4%

[26]Based on a choice-theoretic analysis of local IV estimation, Heckman et al. (2008) show that to estimate the causal effect of one treatment versus another treatment in general requires a covariate (instrument) for each treatment that changes the value of one treatment in the discrete choice model but does not affect both the value of the other treatments and the outcome variable (in our case employment). For the comparison of a treatment to the next best alternative, the necessary assumption is a bit weaker. In this case, only one instrument for the specific treatment investigated is required that satisfies this identification assumption.

for SPST. This shows that local officials not only reacted strongly with respect to sending unemployed individuals into any program, they also targeted specific programs stronger than others. In line with our budget logic, there is a stronger increase in the probability to enter one of the shorter and cheaper measures than for the longer and more expensive ones. This suggests that the management of a LEO aimed at reducing the current relative surplus, but at the same time did not want to bind their hands too much by committing a large share of next year's budget already in advance.

Table 9 displays the second stage estimates for the impact of getting training in one of these programs on later employment status, again separately for year 1, years 2 to 3, and years 4 to 10 after treatment start. Column 1 presents the raw descriptive differences in working between treatment and control group. Starting with the short and medium run, the differences are negative and statistically significant for all four program types STT, PF, SPST, and RT during year 1 and for years 2 to 3 after program start, albeit the association is negligible for SPST during years 2 and 3. The OLS coefficients in column 2, based on the benchmark specification from before which controls for all personal, regional and time variables, are very similar to the descriptive differences. This suggests that the selection with respect to observable characteristics is negligible during the first three years after program start. In other words, caseworkers do not seem to have chosen participants in a way that their observable characteristics are associated systematically with higher or lower employment perspectives.

The LATE coefficients for PF, SPST, and RT imply significantly negative treatment effects, which are stronger than those obtained with OLS. Only for STT, however, they are less negative than the OLS ones and insignificant.[27] These results suggest that the compliers of our budget instrument for PF, SPST, and RT are positively selected on unobservable characteristics with regard to short- and medium-run employment outcomes. Put differently, the lock-in effect is particularly strong for this type of participants. In contrast, the lock-in effect for the compliers for STT is weaker than indicated by the OLS estimates, suggesting a negative selection of unemployed into this program.

Turning to the control function (CF) results provides some further interesting findings.[28] For the longer programs SPST and RT, the CF estimates for year 1 are significantly

[27]Using the same specification as in Table 9, the ATT estimates based on a fractional probit (not reported here) basically coincide with the OLS coefficients. Typically the first two to three digits are the same. We take this as evidence that our specification is sufficiently flexible.

[28]We only report the results for the CF estimates based on specification 3, denoted CF3 above. Wald tests for the significance of the selection correction terms imply that the most flexible specification is to be preferred. In most cases, however, the CF change only little between the three specifications. One

negative but much lower in absolute value than the OLS and the LATE effects. For years 2 to 3, the CF effects for SPST and RT are negative but close to zero and not significant. Thus, the positive selection of the compliers does not carry over to the always takers. To the contrary, the treated for SPST and RT are on average even negatively selected with respect to unobservables and the lock-in effect is basically restricted to year 1. The shorter programs STT and PF, on the other hand, show significantly negative and sizable CF estimates, suggesting a strong negative lock-in effect, which is more pronounced for STT than the associated LATE effect. For PF, the CF coefficients imply a positive selection of the treated with respect to unobservable characteristics, while the selection seems to be negligible compared to all unemployed for STT. This shows that the treated on average do not follow the negative selection of the group of compliers.

We now turn to the long-run effects in years 4 to 10 reported in Table 9. The descriptive differences still show that participants in STT and PF are on average less likely to work in the long run than nonparticipants (by 5.7 and 3.7 ppoints, respectively), while individuals who took part in SPST and RT seem to be more successful by 2.6 and 7.6 ppoints. This pattern of less effective or even harmful effects for the shorter programs, and small, but beneficial results for the longer ones remains present for the OLS estimates. Again the differences are minor, suggesting only negligible selection effects with respect to observable characteristics. Turning to the LATE estimates in column 3, we again find worse effects compared to OLS. While the decline is sizeable for PF, SPST, and RT (with 13, 9, and 6 ppoints, respectively), it is negligible for STT. The LATE itself is only significant for PF, for which it amounts to -14 ppoints. As ATT for all participants, the CF estimates in column 4 suggest better results than OLS in all cases except PF. The results are positive, but insignificant for STT and SPST, whereas RT significantly increases later employment and PF decreases it (by 14 and -21.5 ppoints, respectively). The CF effect for PF is even larger in absolute value than the LATE. Contrasting columns 1 and 4 suggests that with respect to long-run outcomes participants are a negative selection for STT and RT and a positive selection for PF. Selection is negligible for SPST. These results also show that in some cases selection patterns are not invariant over time.

For most cases, our findings imply that the LATE estimates are considerably worse than those obtained from the CF and OLS. Together, this indicates that the individual

notable exception relates to the CF estimate for Long SPST in years 4 to 10, where the CF effect estimate is larger and significant for the less flexible specifications CF1 and CF2. Due to space constraints, these results are omitted here, but can be obtained from the authors.

treatment effect deteriorates when more people are assigned to the program, because the selection regarding observable and unobservable characteristics changes when the program is expanded. This effect is particularly strong for SPST and RT. Thus, we conclude that putting unemployed into training programs as a result of an end-of-year spending effect is not only inefficient, but in most cases even harmful for the involved additional participants, i.e., the compliers of our instrument. More generally, our results demonstrate again that the results of empirical evaluations are sensitive to the choice of the analyzed population. In particular, looking only at the compliers of a certain instrument may lead to different conclusions than taking all participants into account. Also, the effectiveness of the different training programs seems to vary strongly. Specifically, longer programs focusing on the acquisition of new and specific skills fare comparatively better than shorter programs with an emphasis on exercising and practicing more general skills. Above all, the findings of negative long-run effects of practice firms suggest that their existence and format should be closely examined and reconsidered in order to achieve better results for the unemployed with the allocated funds.

As a further test of our conclusions so far, we take a closer look at the different group of measures we have summed up under the label SPST until now. As noted above, these are actually a rather heterogeneous mix of different training courses with varying content and duration, starting from a couple of weeks up to two years. If the story is really short versus long programs, we should also see differences if we distinguish the SPST courses by their duration. Thus, we split them into short and long programs according to whether their planed duration was at least or less than six months, and provide separate estimates. The bottom parts of table 9 present the results of this exercise for each period after program start. We find exactly the pattern as expected from the above considerations. That is, we observe a much smaller lock-in effect during year 1 and a significant negative CF estimate of -9 ppoints in years 4 to 10 for short SPST. In contrast, long SPST measures show a strong and significant negative lock-in effect during year 1, and a positive, but insignificant effect of 7 ppoints in the long run. Thus, even though the CF estimates for long SPST regarding years 4 to 10 are not significant at conventional significance levels, the evidence suggests a pattern of effects over time which is quite similar to the results found for RT. In contrast, there is no evidence suggesting that short SPST is effective.

10 Conclusions

This paper studies the employment effects of training programs for the unemployed up to 10 years after program start. In order to come closer to estimating the causal effect, we take advantage of strict budget rules in the 1980s and early 90s to implement an IV strategy to account for a possible selection on unobservable characteristics of the participants. In particular, we instrument program participation by how much the respective local employment office spent during the first semester of a year relative to its own budget and the other offices. Since a direct transfer of funds from one instrument of active labor market policy to another or into the next year was not allowed, employment offices sitting on comparatively large resources after the summer holidays faced incentives to increase their spending in the months following the summer holidays. We show that such a potential for end-of-year spending led to an increase in the probability of an unemployed individual to enter a training program in the final months of the year which is not related to her observed or unobserved characteristics.

Our empirical analysis of the employment effects of training leads to the following main findings: First, at a methodological level, we show that a flexible control function approach leads to more precise estimates of the program impact compared to standard 2SLS. Second, our estimates for the long-run employment effects of training differ both by the target population we consider (compliers of our instrument vs. average participants) and by the type of program. On the one hand, we see that unemployed who are only selected for a training measure to get rid of budget surpluses do not profit from the program, while the average participant gets a moderate boost in her employment prospects. On the other hand, programs focusing on the acquisition of new specific skills and degrees (retraining) increase the employment of participants in the long run, while shorter programs concentrating on practicing the existing stock of skills (practice firms) even persistently reduce employment prospects. Third, selection based on unobservable characteristics like motivation, ambition, unobserved ability, or strive seems to be strong, but not uniformly positive or negative. While we see evidence for a positive selection of participants in practice firms and short SPST programs, participants in short-term training, long SPST, and retraining seem negatively selected. This suggests that on average these measures are targeted towards individuals with fairly low employment chances. However, the compliers of our budget instrument appear strongly positively selected.

These findings are useful for the policy debate on the effectiveness of training pro-

grams for the unemployed. We find strong evidence for heterogeneity of the employment effects of training. Thus, it is important to analyze the effectiveness of different types of training programs separately. Furthermore, unexpected funding to be spent on training shows particularly bad employment effects. This suggests that caseworkers usually assign individuals with higher potential returns to training first, while additional and unexpected funding leads to a less efficient training assignment. Therefore, local employment offices should have more leeway in their spending instead of having the incentive to exhaust budgets which are exclusively earmarked for training.

References

Aakvik, A., J. J. Heckman, and E. Vytlacil (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics 125*(1–2), 15–51.

Abbring, J. and G. van den Berg (2004). The Nonparametric Identification of Treatment Effects in Duration Models. *Econometrica 71*(5), 1491–1517.

Andersson, F., H. Holzer, J. Lane, D. Rosenblum, and J. Smith (2013). Does Federally-Funded Job Training Work? Nonexperimental Estimates of WIA Training Impacts Using Longitudinal Data on Workers and Firms. *NBER Working Paper 19446*.

Barbe, P. and P. Bertail (1995). *The weighted bootstrap*. Volume 98 of Lecture Notes in Statistics, Springer Publisher, New York.

Bender, S., A. Bergemann, B. Fitzenberger, M. Lechner, R. Miquel, S. Speckesser, and C. Wunsch (2005). Über die Wirksamkeit von FuU-Maßnahmen. *IAB: Beiträge zur Arbeitsmarkt- und Berufsforschung 289*, 410.

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

Blundell, R., L. Dearden, and B. Sianesi (2005). Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society: Series A 168*(3), 473–512.

Blundell, R. and J. Powell (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress II*, ed. by M. Dewatripont, L.P. Hansen, and S.J. Turnovsky. Cambridge, U.K.: Cambridge University Press, 312–357.

Blundell, R. and J. Powell (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies 71*(3), 655–679.

Bundesanstalt für Arbeit (1993). Geschäftsbereicht der Bundesanstalt für Arbeit (ANBA).

Caliendo, M., R. Mahlstedt, and O. Mitnik (2014). Unobservable, but Unimportant? The Influence of Personality Traits (and Other Usually Unobserved Variables) for the Evaluation of Labor Market Policies. *IZA Discussion Paper 8337*.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Card, D., J. Kluve, and A. Weber (2010). Active labor market policy evaluations: A meta-analysis. *The Economic Journal 120*(4), 742–784.

Card, D., J. Kluve, and A. Weber (2015). What works? a meta-analysis of recent active labour market evaluations. Technical Report 9236, IZA Discussion Paper.

Chiburis, R., J. Das, and M. Lokshin (2012). A practical comparison of the bivariate probit and linear IV estimators. *Economics Letters 117*(3), 762–766.

Comptroller General (1980). Federal Year-End Spending - Symptom of a larger problem. *US House of Representatives*.

Douglas, J. W. and A. L. Franklin (2006). Putting the Brakes on the Rush to Spend Down End-of-Year Balances: Carryover Money in Oklahoma State Agencies. *Public Budgeting & Finance 26*(3), 46–64.

Fertig, M. and C. Schmidt (2000). Discretionary measures of active labor market policy: the German employment promotion reform in perspective. *IAB Discussion Paper 182*.

Fitzenberger, B. and G. Muehler (2015). Dips and floors in workplace training: gender differences and supervisors. *Scottish Journal of Political Economy 62*(4), 400–429.

Fitzenberger, B., O. Orlanski, A. Osikominu, and M. Paul (2013). Déjà Vu? Short-Term Training in Germany 1980-1992 and 2000-2003. *Empirical Economics 44*(1), 289–328.

Fitzenberger, B., A. Osikominu, and M. Paul (2010). The heterogeneous effects of training incidence and duration on labor market transitions. *IZA Discussion Paper 5269*.

Fitzenberger, B., A. Osikominu, and R. Völter (2008). Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany. *Annales d'Economie et de Statisque 91–92*, 321–355.

Fitzenberger, B. and S. Speckesser (2007). Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany. *Empirical Economics 32*(2/3), 529–573.

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

Frölich, M. and M. Lechner (2010). Exploiting regional treatment intensity for the evaluation of labour market policies. *Journal of the American Statistical Association 105*(491), 1014–1029.

General Accounting Office (1998). Year-end spending: Reforms underway but better reporting and oversight needed. *Publication No. GAO/AIMD–98–185 Washington, D.C.: U.S. Government Printing Office*.

Ham, J. C. and R. J. LaLonde (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica*, 175–205.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*(4), 931–959.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heckman, J. J., R. LaLonde, and J. Smith (1999). The Economics and Econometrics of Active Labor Market Programs. *In: O. Aschenfelter and D. Card (eds.), Handbook of Labor Economics Vol. 3 A, Amsterdam: Elsevier Science*, 1865–2097.

Heckman, J. J., S. Urzua, and E. Vytlacil (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique*, 151–174.

Heinrich, C., P. Mueser, K. Troske, K. Jeon, and D. Kahvecioglu (2013). Do Public Employment and Training Programs Work? *IZA Journal of Labor Economics 2:6*, 1–23.

Kimhi, A. (1999). Estimation of an endogenous switching regression model with discrete dependent variables: Monte-Carlo analysis and empirical application of three estimators. *Empirical Economics 24*(2), 225–241.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour Economics 17*(6), 904–917.

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

Lee, L.-F. (1982). Some Approaches to the correction of selectivity bias. *The Review of Economic Studies 49*(3), 355–377.

Li, Y. P., K. J. Propert, and P. R. Rosenbaum (2001). Balanced risk set matching. *Journal of the American Statistical Association 96*(455), 870–882.

Liebman, J. and N. Mahoney (2013). Do expiring budgets lead to wasteful year-end spending? Evidence from federal procurement. *NBER Working Paper 19481.*

Martin, J. P. (2015). Activation and active labour market policies in oecd countries: stylised facts and evidence on their effectiveness. *IZA Journal of Labor Policy 4*(1), 1.

McPherson, M. (2007). An analysis of year-end spending and the feasibility of a carryover incentive for federal agencies. *Master dissertation. Naval Postgraduate School.*

Merchant, K. A. (1985). Budgeting and the propensity to create budgetary slack. *Accounting, Organizations and Society 10*(2), 201–210.

Osikominu, A. (2013). Quick job entry or long-tern human capital development? The dynamic effects of alternative training schemes. *Review of Economic Studies 80*(1), 313–342.

Richardson, K. and G. J. van den Berg (2013). Duration dependence versus unobserved heterogeneity in treatment effects: Swedish labor market training and the transition rate to employment. *Journal of Applied Econometrics 28*(2), 325–351.

Rivers, D. and Q. Vuong (1988). Limited information estimators and exogeneity test for simultaneous probit models. *Journal of Econometrics 39*(3), 347–366.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Rubin, D. (1974). Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology 66*(5), 688–701.

Schneider (1981). Erfahrungen mit "41a". *Arbeit und Beruf 4*(1981), 97–99.

Sianesi, B. (2004). An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s. *Review of Economics and Statistics 86*(1), 133–155.

Sianesi, B. (2008). Differential effects of active labour market programs for the unemployed. *Labour Economics 15*(3), 370–399.

Solon, G., S. J. Haider, and J. M. Wooldridge (2015). What are we weighting for? *Journal of Human resources 50*(2), 301–316.

Terza, J. (2009). Parametric nonlinear regression with endogeneous switching. *Econometric Reviews 28*(6), 555–580.

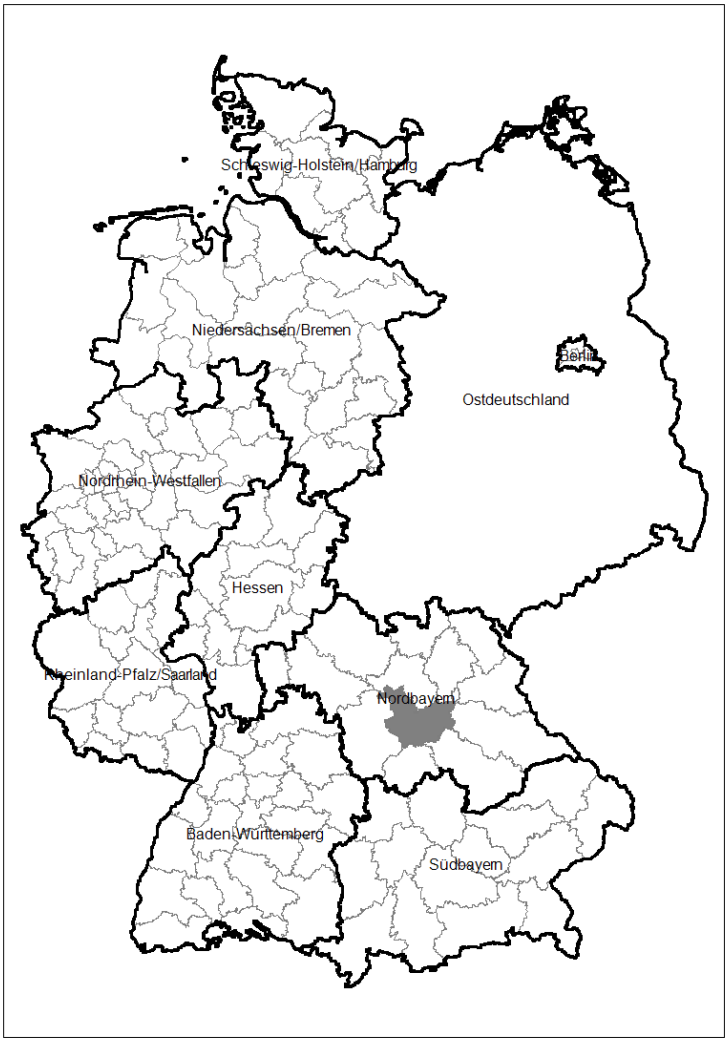
Terza, J., A. Basu, and P. Rathouz (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics 27*(3), 531–543.

Wooldridge, J. (2005). Unobserved heterogeneity and estimation of average partial effects. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D.W.K. Andrews and J.H. Stock. Cambridge, U.K.: Cambridge University Press, 27–55.

Wooldridge, J. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics 182*(1), 226–234.

Figures and tables

Figure 1: Organizational structure of the Federal Employment Office (1980s)



Note: Lines in bold black mark the borders of the nine regional employment offices, those in light grey the limits of the 142 local employment offices. The location of the federal headquarter in Nuremberg is indicated by the corresponding LEO appearing in solid grey.

Figure 2: The employment office's budget year

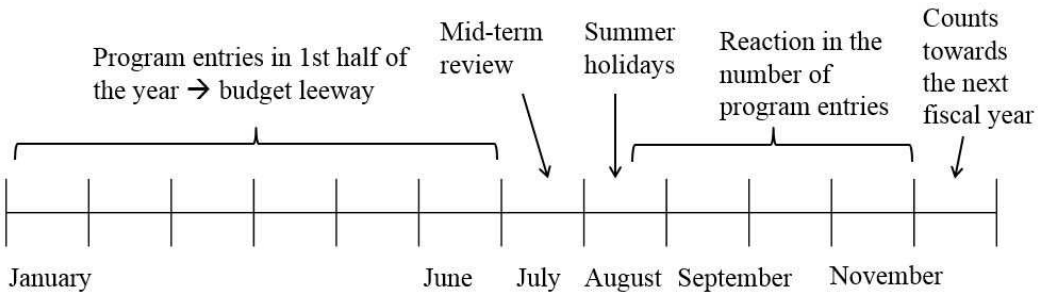
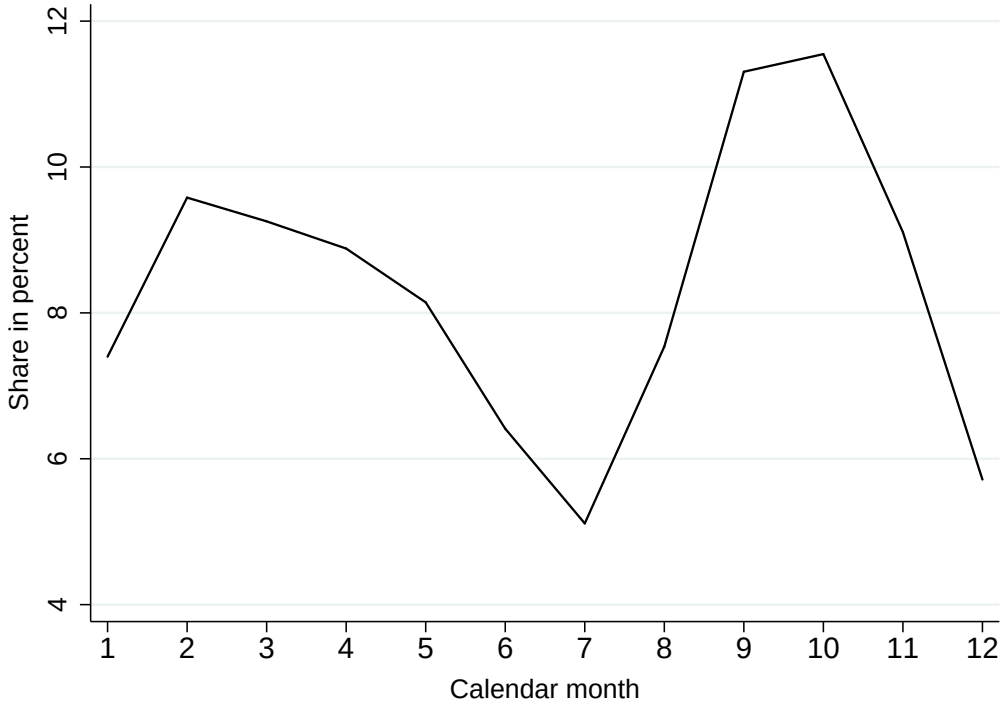
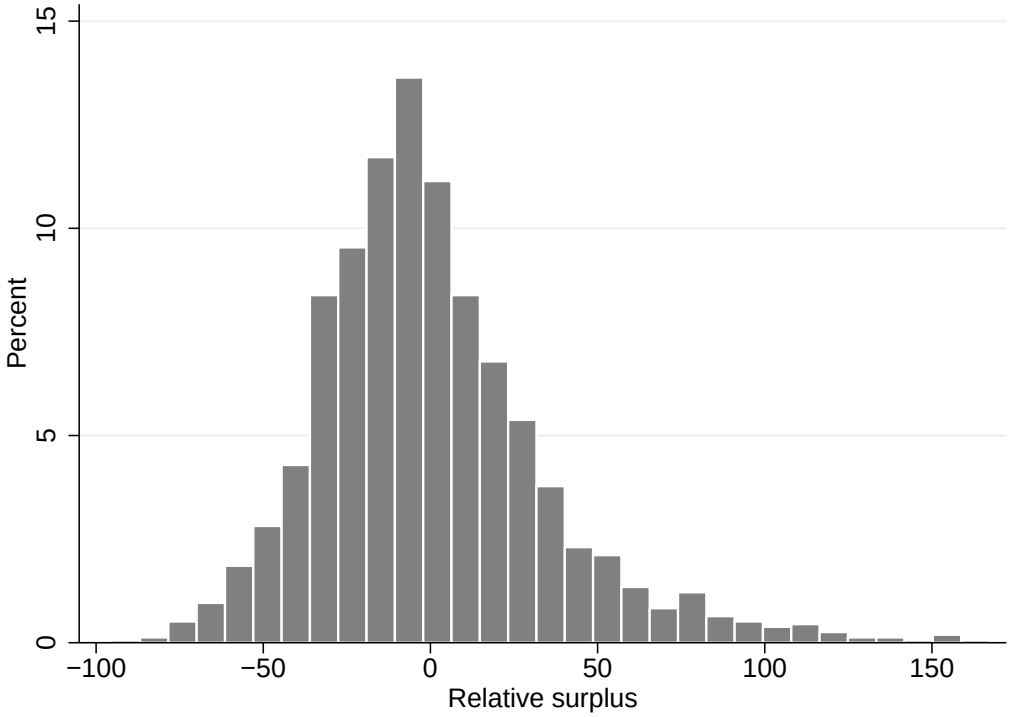


Figure 3: Monthly shares of total year entries into training programs, 1983 – 1993



Note: Shares are calculated on the basis of entry numbers averaged over region and time. The computation is based upon the 50% sample of all participants.

Figure 4: Distribution of relative surplus



Note: Relative surplus is defined as the difference between a local employment office's budget leeway after the first half of the year and the average budget leeway of all other districts in that year. Budget leeway is calculated as the difference between planned budget for the entire year and actual entries over the first six months of the year, normalized by the number of eligible unemployed individuals (in 1,000) in July. The calculations use population weights.

Table 1: Size of treatment and control group

Elapsed duration	Control group	Treatment group	STT[a)]	PF	SPST	RT
1–3	771,139	36,782	3,813	3,070	19,446	10,453
4–6	426,544	24,261	3,186	2,327	13,017	5,731
7–9	320,852	19,214	2,698	2,040	9,535	4,941
10–12	240,373	14,679	2,358	1,670	7,057	3,594
1–12	1,758,908	94,936	12,055	9,070	49,055	24,719

Note: [a)] Since STT programs existed only until the end of 1992, the control group for this training is a bit smaller and amounts to 1,470,480 nontreated observations. Absolute numbers based upon 50% sample of all treated and 3% sample of all nontreated.

Table 2: Summary statistics for selected explanatory variables

	Control group		Treatment group		STT		PF		SPST		RT	
Variable	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev
Female and age												
Female	0.400	(0.490)	**0.449**	(0.497)	**0.445**	(0.497)	**0.329**	(0.500)	**0.482**	(0.500)	**0.431**	(0.495)
20-29 years old	0.279	(0.449)	**0.320**	(0.467)	**0.302**	(0.459)	**0.299**	(0.458)	**0.280**	(0.449)	**0.417**	(0.493)
30-34 years old	0.224	(0.417)	**0.249**	(0.432)	0.227	(0.419)	0.225	(0.418)	**0.240**	(0.427)	**0.285**	(0.452)
35-39 years old	0.174	(0.379)	**0.180**	(0.384)	**0.180**	(0.384)	**0.185**	(0.388)	**0.189**	(0.391)	**0.162**	(0.368)
40-44 years old	0.153	(0.360)	**0.139**	(0.346)	0.148	(0.550)	0.154	(0.361)	**0.159**	(0.365)	**0.091**	(0.288)
45-50 years old	0.171	(0.376)	**0.112**	(0.315)	**0.144**	(0.351)	**0.137**	(0.343)	**0.133**	(0.340)	**0.045**	(0.207)
Education												
No voc. training degree	0.195	(0.396)	**0.159**	(0.366)	**0.216**	(0.411)	**0.207**	(0.405)	**0.118**	(0.323)	0.197	(0.398)
Voc. Training degree	0.718	(0.450)	**0.753**	(0.431)	0.723	(0.447)	**0.767**	(0.423)	**0.777**	(0.416)	0.714	(0.452)
Uni/college degree	0.083	(0.276)	0.084	(0.278)	**0.056**	(0.230)	**0.023**	(0.149)	**0.102**	(0.302)	0.086	(0.281)
Education unknown	0.005	(0.070)	**0.003**	(0.054)	0.005	(0.067)	**0.003**	(0.054)	**0.003**	(0.052)	**0.003**	(0.051)
Marital stataus, children in household, foreigner												
Married	0.474	(0.499)	**0.449**	(0.497)	**0.445**	(0.497)	**0.446**	(0.497)	0.475	(0.499)	**0.399**	(0.490)
Kids	0.341	(0.474)	**0.336**	(0.472)	**0.304**	(0.460)	0.350	(0.477)	**0.349**	(0.477)	**0.322**	(0.467)
Foreigner	0.118	(0.322)	**0.071**	(0.258)	**0.077**	(0.267)	**0.093**	(0.291)	**0.066**	(0.248)	**0.072**	(0.258)
Previous employment and employment stataus												
# months employed	18.812	(6.286)	18.881	(6.367)	18.536	(6.535)	**18.114**	(6.628)	**19.187**	(6.279)	18.723	(6.320)
Log wage	3.794	(0.799)	**3.833**	(0.748)	**3.814**	(0.666)	**3.854**	(0.614)	**3.834**	(0.799)	**3.833**	(0.726)
Apprentice	0.015	(0.122)	0.014	(0.118)	**0.011**	(0.102)	**0.011**	(0.103)	**0.013**	(0.113)	**0.019**	(0.138)
Blue collar worker	0.570	(0.495)	**0.474**	(0.499)	**0.559**	(0.497)	**0.634**	(0.482)	**0.381**	(0.486)	**0.559**	(0.496)

<continued on next page>

Table 2 – <continued from previous page>

	Control group		Treatment group		STT		PF		SPST		RT	
Variable	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev
White collar worker	0.316	(0.465)	**0.416**	(0.493)	**0.343**	(0.475)	**0.286**	(0.452)	**0.501**	(0.500)	**0.329**	(0.470)
Worker at home	0.003	(0.058)	**0.001**	(0.034)	**0.001**	(0.036)	**0.001**	(0.031)	**0.001**	(0.031)	**0.002**	(0.039)
Part-time worker	0.095	(0.294)	0.095	(0.293)	**0.087**	(0.281)	**0.068**	(0.252)	**0.104**	(0.306)	0.090	(0.287)
Occupation from previous employment												
Farmer/Fisher	0.028	(0.165)	**0.021**	(0.144)	0.028	(0.166)	**0.039**	(0.193)	**0.016**	(0.124)	**0.022**	(0.147)
Manufacturing	0.373	(0.484)	**0.309**	(0.462)	**0.357**	(0.479)	**0.416**	(0.493)	**0.261**	(0.439)	**0.341**	(0.474)
Technicians	0.039	(0.193)	**0.051**	(0.219)	**0.030**	(0.171)	**0.018**	(0.133)	**0.072**	(0.258)	**0.031**	(0.173)
Service	0.532	(0.499)	**0.597**	(0.490)	**0.555**	(0.497)	**0.476**	(0.499)	**0.636**	(0.481)	**0.586**	(0.493)
Miners/Others/Missing	0.028	(0.165)	**0.022**	(0.147)	0.029	(0.168)	**0.052**	(0.221)	**0.016**	(0.124)	**0.020**	(0.141)
Firm size from prvious employment												
$<$ 10 employees	0.253	(0.435)	**0.232**	(0.422)	**0.226**	(0.418)	**0.219**	(0.413)	**0.240**	(0.427)	**0.225**	(0.418)
$\geq$ 10 & $<$ 20 employees	0.118	(0.322)	**0.113**	(0.316)	**0.111**	(0.314)	0.119	(0.323)	0.116	(0.320)	**0.106**	(0.308)
$\geq$ 20 & $<$ 50 employees	0.152	(0.359)	0.152	(0.359)	0.147	(0.354)	**0.160**	(0.367)	0.153	(0.360)	0.149	(0.356)
$\geq$ 50 &200 employees	0.202	(0.401)	**0.214**	(0.410)	**0.214**	(0.410)	**0.227**	(0.419)	**0.211**	(0.408)	**0.216**	(0.411)
$\geq$ 200 & $<$ 500 employees	0.100	(0.300)	**0.109**	(0.312)	**0.111**	(0.314)	**0.113**	(0.316)	**0.107**	(0.309)	**0.111**	(0.314)
$\geq$ 500 employees	0.140	(0.347)	**0.146**	(0.540)	**0.156**	(0.363)	0.131	(0.338)	0.139	(0.346)	**0.163**	(0.369)
Missing	0.035	(0.184)	0.033	(0.179)	0.034	(0.182)	0.033	(0.178)	0.035	(0.183)	**0.030**	(0.171)
Former treatment participation												
1 year before	0.036	(0.185)	**0.049**	(0.216)	**0.042**	(0.201)	**0.066**	(0.248)	**0.053**	(0.225)	**0.039**	(0.193)
2 years before	0.060	(0.238)	**0.075**	(0.264)	**0.067**	(0.250)	**0.103**	(0.304)	**0.081**	(0.273)	0.057	(0.233)

Note: Mean and standard deviation for selected explanatory variables are reported over the first 12 months of unemployment. Bold cases indicate significant difference in the mean value between treatment (in the respective program) and control group at 1% or 5% significance level. Calculations are based on weighted individual observations where each individual is replicated based on the number of months he/she remains unemployed and is eligible for treatment. All calculations are based upon the 50% sample of all treated and the 3% sample of all non-treated.

Table 3: The effect of relative surplus on the probability to enter any program (1st stage)

	(1)	(2)	(3)	(4)	(5)
Predictions on planned spending (last three years)					
Average partial effect	2.577***	2.194***	1.922***	1.883***	1.866***
Standard error	(0.178)	(0.231)	(0.218)	(0.223)	(0.227)
F-statistic	209.671	90.083	78.097	71.186	67.684
Predictions on planned spending (last two years)					
Average partial effect	2.613***	2.118***	1.846***	1.817***	1.806***
Standard error	(0.176)	(0.221)	(0.203)	(0.210)	(0.213)
F-statistic	220.621	91.949	82.589	75.097	71.745
Predictions on planned spending (last four years)					
Average partial effect	2.308***	1.968***	1.696***	1.649***	1.617***
Standard error	(0.184)	(0.251)	(0.235)	(0.242)	(0.246)
F-statistic	157.886	61.610	52.009	46.384	43.064
Region & time info[a)]	yes	yes	yes	yes	yes
Personal characteristics[b)]	yes	yes	yes	yes	yes
Relative surplus in $\tau - 1$	no	yes	yes	yes	yes
Region $\times$ year interactions	no	no	yes	yes	yes
Local UR (Jan-July)[c)]	no	no	no	yes	yes
Local UR (last year)[d)]	no	no	no	no	yes

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. The numbers report the effect of an increase in relative surplus by one standard deviation in ppoints on the probability to enter any program for separate OLS regressions. Clustered standard errors at local labor market level and time are reported in parentheses, the corresponding F-statistics in square brackets. [a)]Region and year fixed effects, calendar month dummies, share of summer holidays in year and region, and interactions between share of summer holidays, calendar month, and year. [b)]Individual characteristics, information on previous employment, former treatment participation, and elapsed unemployment duration. [c)]Local unemployment rate (LUR) in each month from January to July in the respective year of program start. [d)]LUR in each month of the year $\tau - 1$ (τ is the year of treatment start). Calculations based upon our full sample of all treated and a 20% random sample of the available nontreated. When used in the comparison group, individuals treated in the future are weighted down relative to nontreated individuals.

Table 4: Effect of training on subsequent employment of compliers (LATE)

Year	(1) Desc	(2) OLS	(3) OLS	(4) Reduced Form	(5) Reduced Form	(6) LATE	(7) LATE
1	-0.192***	-0.198***	-0.198***	-0.003**	-0.003*	-0.171**	-0.138*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.078)	(0.080)
2–3	-0.054***	-0.055***	-0.055***	-0.002	-0.002	-0.132	-0.103
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.104)	(0.106)
4–10	0.023***	0.021***	0.021***	-0.002	-0.001	-0.083	-0.03
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.096)	(0.098)
Controls	no	yes	yes	yes	yes	yes	yes
Local UR (last year)	no	no	yes	no	yes	no	yes

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. Standard errors (in parentheses) are clustered at time and local labor market level and obtained through weighted bootstrapping based on 200 replications. Controls include region and time information, personal characteristics, first lag of relative surplus, interaction terms between region and year, time path (January-July) of local unemployment rate in the year of treatment start. Local UR (last year) represents the local unemployment rate in each month of the previous year. Sample as in Table 3.

Table 5: Specification tests for including the generalized residuals

Year	1	2–3	4–10
		CF 1	
ω_0	0.039 (0.091)	-0.030 (0.090)	-0.251 (0.077)***
ω_1	0.143 (0.040)***	-0.009 (0.034)	0.021 (0.031)
$H0 : \omega_0 = \omega_1 = 0$	20.794 [0.000]	0.226 [0.893]	23.368 [0.000]
		CF 2	
ω_0	0.030 (0.088)	-0.066 (0.091)	-0.239 (0.077)***
ω_1	0.011 (0.058)	0.107 (0.041)***	0.132 (0.041)***
$H0 : \omega_0 = \omega_1 = 0$	0.118 [0.943]	12.256 [0.002]	30.996 [0.000]
		CF 3	
ω_0	0.274 (0.152)*	0.051 (0.175)	0.049 (0.163)
ω_1	-0.040 (0.056)	0.099 (0.040)**	0.148 (0.040)***
$H0 : \omega_0 = \omega_1 = 0$	5.548 [0.062]	6.220 [0.045]	13.948 [0.001]
$H0 : \psi = 0$	660.442 [0.000]	365.454 [0.000]	336.001 [0.000]

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. Standard errors (in parentheses), test statistics (χ^2-statistic from a Wald test of joint significance of parameters), and p-values (in brackets) are clustered at time and local labor market level and obtained through weighted bootstrapping based on 200 replications. Benchmark specification controls for regional and time variation, individual characteristics, lagged value of relative surplus, time path (January-July) of local unemployment rate in the year of treatment start, and local unemployment rate from the previous year. CF 1 includes generalized residuals and interaction between generalized residuals and treatment dummy. CF 2 includes in addition to CF 1 interaction terms between individual characteristics and treatment dummy. CF 3 includes in addition to CF 2 interaction terms between generalized residuals and individual characteristics. Sample as in Table 3.

Table 6: ATT of Training on Subsequent Employment - Specification 4

Year	OLS	FP	LATE	CF 1	CF 2	CF 3
1	-0.198***	-0.199***	-0.171**	-0.240***	-0.236***	-0.183***
	(0.002)	(0.002)	(0.078)	(0.057)	(0.056)	(0.051)
2–3	-0.055***	-0.055***	-0.132	-0.049	-0.027	-0.028
	(0.002)	(0.002)	(0.104)	(0.057)	(0.059)	(0.056)
4–10	0.021***	0.021***	-0.083	0.156***	0.149***	0.101**
	(0.002)	(0.002)	(0.096)	(0.048)	(0.048)	(0.045)

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. Standard errors (in parentheses) are clustered at time and local labor market level and obtained through weighted bootstrapping based on 200 replications. Specification (4) controls for regional and time variation, individual characteristics, lagged value of relative surplus, and time path (January-July) of local unemployment rate in the year of treatment start. Sample as in Table 3.

Table 7: ATT of Training on Subsequent Employment - Specification 5

Year	OLS	FP	LATE	CF 1	CF 2	CF 3
1	-0.198***	-0.199***	-0.138*	-0.225***	-0.218***	-0.165***
	(0.002)	(0.002)	(0.080)	(0.058)	(0.057)	(0.051)
2–3	-0.055***	-0.055***	-0.103	-0.037	-0.014	-0.016
	(0.002)	(0.002)	(0.106)	(0.058)	(0.059)	(0.056)
4–10	0.021***	0.021***	-0.030	0.180***	0.172***	0.124***
	(0.002)	(0.002)	(0.098)	(0.048)	(0.048)	(0.045)

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. Standard errors (in parentheses) are clustered at time and local labor market level and obtained through weighted bootstrapping based on 200 replications. Specification (5) controls for regional and time variation, individual characteristics, lagged value of relative surplus, time path (January-July) of local unemployment rate in the year of treatment start, and local unemployment rate from the previous year. Sample as in Table 3.

Table 8: The effect of relative surplus on treatment probability in different training programs (OLS results)

	(1) STT	(2) PF	(3) SPST	(4) RT	(5) SSPST	(6) LSPST
Average partial effect	2.527***	2.715***	1.695***	1.120**	1.457***	1.629***
Standard error	(0.665)	(0.419)	(0.330)	(0.359)	(0.391)	(0.398)
F-statistic	14.415	41.975	26.346	9.738	13.920	16.716
Monthly hazard rate (in %)	22.32	15.59	49.86	33.39	37.16	28.74

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. The numbers report the effect of an increase of one standard deviation in relative surplus on the probability to enter the respective training program in ppoints. STT: Short-term training; PF: Practice firms; SPST: Specific professional skills and techniques; RT: Retraining; SSPST: Short SPST (planned duration <= 6 months); LSPST: Long SPST (planned duration > 6 months). Clustered standard errors at local labor market level and time are reported in parentheses, the corresponding F-statistics in square brackets. Benchmark specification: controls for regional and time variation, individual characteristics, lagged value of relative surplus, time path (January-July) of local unemployment rate in the year of treatment start, and local unemployment rate from the previous year. Specification corresponds to specification in column 5 of Table 3. Samples as in Table 3, excluding participants in other training programs.

Table 9: Employment effects of training (by type of program)

Training program	(1) Des	(2) OLS	(3) LATE	(4) CF
		Year 1		
Short-term training (STT)	-0.168***	-0.151***	-0.052	-0.147***
	(0.003)	(0.003)	(0.058)	(0.057)
Practice firms (PF)	-0.165***	-0.146***	-0.187***	-0.192***
	(0.004)	(0.002)	(0.063)	(0.058)
Specific prof. skills	-0.146***	-0.159***	-0.185*	-0.107**
and techniques (SPST)	(0.002)	(0.002)	(0.101)	(0.048)
Retraining (RT)	-0.308***	-0.316***	-0.400***	-0.122**
	(0.002)	(0.002)	(0.058)	(0.050)
Short SPST	-0.092***	-0.108***	-0.059	-0.033
	(0.003)	(0.003)	(0.126)	(0.051)
Long SPST	-0.225***	-0.236***	-0.299***	-0.261***
	(0.002)	(0.002)	(0.101)	(0.049)
		Years 2-3		
Short-term training (STT)	-0.120***	-0.098***	-0.041	-0.119*
	(0.005)	(0.004)	(0.058)	(0.064)
Practice firms (PF)	-0.074***	-0.048***	-0.130*	-0.098*
	(0.005)	(0.004)	(0.075)	(0.054)
Specific prof. skills	-0.006*	-0.008***	-0.234*	-0.008
and techniques (SPST)	(0.003)	(0.003)	(0.141)	(0.056)
Retraining (RT)	-0.136***	-0.129***	-0.217***	-0.032
	(0.003)	(0.003)	(0.058)	(0.061)
Short SPST	0.014***	-0.007**	-0.164	-0.066
	(0.003)	(0.003)	(0.160)	(0.059)
Long SPST	-0.006	-0.012***	-0.308**	-0.019
	(0.004)	(0.003)	(0.141)	(0.052)

<continued on next page>

Table 9 – <continued from previous page>

	(1) Des	(2) OLS	(3) LATE	(4) CF
Years 4-10				
Short-term training (STT)	-0.057***	-0.045***	-0.063	0.017
	(0.005)	(0.004)	(0.048)	(0.067)
Practice firms (PF)	-0.037***	-0.011***	-0.143**	-0.215***
	(0.005)	(0.004)	(0.073)	(0.047)
Specific prof. skills	0.026***	0.013***	-0.076	0.025
and techniques (SPST)	(0.003)	(0.002)	(0.117)	(0.043)
Retraining (RT)	0.076***	0.076***	0.017	0.140**
	(0.003)	(0.003)	(0.048)	(0.066)
Short SPST	0.023***	0.003	-0.101	-0.094*
	(0.003)	(0.002)	(0.144)	(0.055)
Long SPST	0.030***	0.026***	-0.154	0.069
	(0.004)	(0.003)	(0.117)	(0.048)

Note: ***, **, and * indicate statistical significance at 1%, 5%, and 10% level, respectively. Standard errors (in parentheses) are clustered at time and local labor market level and obtained through weighted bootstrapping based on 200 replications. Benchmark specification: controls for regional and time variation, individual characteristics, lagged value of relative surplus, time path (January-July) of local unemployment rate in the year of treatment start, and local unemployment rate from the previous year. Specification corresponds to specification CF 3 in Table 7. Samples as in Table 3, excluding participants in other training programs.

A Appendix

A.1 Details on the Construction of the Data Set

All samples used in this study were drawn according to the so called "birthday concept". That is, in the samples both from the FuU data and from the IEB data, 50% of all possible birthdays starting with January, 2nd, are drawn and all observations with those 182 birthdays included. The 3% IEB sample was obtained in the same way, just that here only 12 of the 182 birthdays chosen above are considered and all records that have already been drawn before are dropped.

The combined raw data had a spell form and contained a lot of temporal overlaps. We carried out a number of corrections, mostly based on Bender et al. (2005), in order to improve data quality and prepare the data for the empirical analysis. The most important data preparation steps involved extending the FuU data with information from IEB. The merge procedure was based on a personal identification number and additional criteria like consistency in time structure and contents of the corresponding spells. For all data sources, we adjusted the temporal overlaps between the different types of spells, corrected the education variable according to imputation rules developed by Fitzenberger et al. (2006), and generated the data on a monthly basis.

The calculation of average outcomes based on the raw data for aggregates at the local level, e.g. the local unemployment rate, uses all treated and controls applying the sampling weights 12 versus 182. Because the 3% control sample is still very large regarding the absolute number of individuals compared to the treatment sample, the relative weight of a treated individual is five times the weight of an individual in the control sample for the estimation of treatment effects (and for the descriptive differences between treated and controls), see also footnote 19. For the empirical analysis using risk set matching, we weight the observations depending on the data source and on the treatment status based on elapsed unemployment duration. Individuals who are treated later on receive a lower weight to compensate for the oversampling of the treated when they serve as controls before they are treated. Thus, controls who never participate in a training program (i.e., those from the 3% IEB sample) receive a weight of $5 * 182/12$ relative to those controls who later participate in training.

A.2 Control Function Approach to Estimate ATT

To estimate a random coefficients model with both a binary endogenous treatment and a fractional outcome variable, we adopt a flexible CF approach for nonlinear models with discrete explanatory endogenous variables as described in Wooldridge (2014, section 6).[29] The idea is that a control function derived from a variable addition test for endogeneity can be used in a flexible way in a one-step or two-step quasi-maximum likelihood framework to identify and estimate treatment effects based on the estimation of the average structural function as introduced by Blundell and Powell (2003).[30] Both for computational simplicity and for the possibility to use a vector of flexible control functions, we opt for the two-step control function approach to estimate the average structural function.[31] Wooldridge (2014) shows how the estimated average structural function can be used to estimate the average treatment effect (ATE). We extend the approach to estimate the average effect of treatment on the treated (ATT).

Following Wooldridge (2014), we maintain the following assumptions for identification:

(A1) $E\left[y_t \mid z, d, b_t^d, u_t\right] = E\left\{\mathbb{1}\left[a_{t0} + z_1 b_0 + (b_{t0} + z_1(b_1 - b_0))d + b_t^d d + u_t \geq 0\right] \mid z, d, b_t^d, u_t\right\}$,

(A2) $E\left[y_t \mid d, z_1, b_t^d, u_t, e_2\right] = E\left[y_t \mid d, z_1, b_t^d, u_t\right]$, and

(A3) $D(b_t^d, u_t \mid z, d) = D(b_t^d, u_t \mid e_2)$.

The outcome variable of interest, $y_t = \sum_{j \in M_t} y_j / m_t$ [with y_j is the employment dummy in month j, M_t represents the different months j in time period t, and m_t the length of period t in months] is the fractional employment rate in period t since treatment start (the average of monthly employment dummies). Following the potential outcome framework (Roy, 1951; Rubin, 1974), we assume that for each individual there are two potential outcomes, $\{y_t^1, y_t^0\}$ at time t associated with a time invariant binary treatment indicator d determined at the time of treatment start, which is elapsed month of unemployment el

[29] The approach builds on Wooldridge (2005) and Terza et al. (2008) as well as on earlier work by Heckman (1978), Lee (1982), Rivers and Vuong (1988), and Blundell and Powell (2003).

[30] Similarly, Terza et al. (2008) suggest a computationally simple "two-stage residual inclusion" approach in a parametric nonlinear regression framework where the residuals from a first stage regression for an endogenous treatment dummy can be used as a control function. Terza (2009) suggests a computationally more expensive estimation approach which relies on correctly integrating out the control function, i.e. the distribution of the unobserved heterogeneity term given the endogenous treatment dummy, in a nonlinear regression specification.

[31] Wooldridge (2014, p. 233) points out that the two-step control function approach involves a different parametric approximation compared to one-step bivariate probit estimation, which tightly specifies the joint distribution of the error terms. Similarly, Terza's (2008) two-stage residual inclusion approach involves yet another parametric approximation.

in risk set matching. $d = 1$ indicates participation in training and $d = 0$ nonparticipation. We assume further that the observed outcome variable is expressed in terms of potential outcomes as $y_t = y_t^0 + (y_t^1 - y_t^0)d$. We impose the following latent index structure for y_t and d:

$$\text{(5a)} \quad E(y_t|d, z_1) = Pr\left[y_{t,j}^* \geq 0\right] = Pr\left[a_{t0} + z_1 b_0 + (b_{t0} + z_1(b_1 - b_0))d + b_t^d d + u_t \geq 0\right]$$

$$\text{(5b)} \quad d = \mathbb{1}\left[d^* \geq 0\right] = \mathbb{1}\left[\gamma_0 + z_1\gamma_1 + z_2\gamma_2 + \nu \geq 0\right] = \mathbb{1}\left[z\gamma + \nu \geq 0\right],$$

$y_{t,j}^*$ and d^* (both defined implicitely) are latent indices, z_1 involves the observed exogenous covariates, z_2 is the set of excluded instruments, $z \equiv (z_1, z_2)$, and b_t^d, u_t and ν are unobserved random variables. $\mathbb{1}[\mathbb{A}]$ denotes the indicator function with a value of one if $\mathbb{A}$ is true and of zero otherwise. We allow for separate time effects (a_{t0}, b_{t0}) and for separate effects (b_0, b_1) of the covariates z_1 by treatment status. We assume a probit model for the treatment dummy, i.e. $\nu \mid z \sim \mathcal{N}(0, 1)$. The potential outcome representation in equation (5a) accounts for selection into treatment based on observable characteristics (z_1, z_2), unobservables u_t, and unobservable random gains from treatment b_t^d. We assume $E(b_t^d \mid z) = 0$. Selection on unobservables is reflected in the statistical dependency between ν and (b_t^d, u_t) (Wooldridge (2014), section 6.1; Blundell et al. (2005), section 3.4.1). Furthermore, we assume that b_t^d and u_t each follow a univariate normal distribution and that (b_t^d, u_t, ν) follow a joint continuous distribution conditional on z. The conditional distribution of (b_t^d, u_t) given ν is allowed to depend upon z_1.

The variables z_1 involve information on (i) individual characteristics like gender, age, education, family status, nationality; (ii) occupation- and job-related variables from previous employment like employment status, earnings, firm size, and industry structure; (iii) individual work history and indicators of former participation; (iv) regional information at the state level as well as time specific variables. The instrument z_2 involves the budget surplus variable.

Assumption **(A1)** specifies the structural expectation as a fractional probit response function with scaled coefficients (Wooldridge, 2005, 2014). Without further assumptions, treatment effects are not identified from the conditional expectation function, because the outcome variable y_t does not only depend upon observed characteristics but also on the unobserved heterogeneity effects, (b_t^d, u_t), which we allow to depend upon the treatment variable via ν.

Assumption **(A2)** is an ignorability condition on the control functions e_2 in the struc-

tural conditional expectation and essentially holds by the definition of e_2. It means that once observed and unobserved factors are controlled for in the response function, proxies for observed and unobserved heterogeneity are redundant for y_t. Under the assumption that selection into treatment can be described by a probit model, a natural choice for the control function is using the generalized residual gr of the probit model (gr involves the standard Heckman (1978, 1979) selection correction term $gr = d\lambda(z\gamma) - (1-d)\lambda(-z\gamma)$, where $\lambda(.)$ denotes the inverse Mills' ratio and γ the first stage probit coefficients. Wooldridge (2014) shows that under correct specification of the probit model for d, a variable addition test for treatment exogeneity based on the generalized residuals is asymptotically optimal. Note that we estimate the effect of treatment started at some point of time on future outcomes in period t, thus gr is determined at treatment start and does not change over time t. Nevertheless, the control function e_2 may vary over time because (b_t^d, u_t) can change over time.

Assumption **(A3)** imposes ignorability restrictions on the conditional distribution of unobserved heterogeneity, such that conditioning on e_2 in the structural expectation is sufficient to correct for selectivity bias arising from the endogeneity of treatment (Wooldridge, 2014).[32] Since the endogenous regressor in our application is a dummy variable, we should view the ignorability assumption about the conditional distribution of (b_t^d, u_t) only as an approximation for a given vector of control functions. Note that the impact of the selection correction term is not nonparametrically identified because the sign of the generalized residual gr is perfectly collinear with the treatment dummy (see Wooldridge, 2014, section 6.3). This is in contrast to the case of a continuous endogenous regressor as discussed in Blundell and Powell (2003). To increase the flexibility and the robustness of the analysis, Wooldridge (2014) suggests adding the square of gr, interactions between gr and the treatment dummy and between gr and the observed characteristics z_1 to the vector of control functions.[33]

Under assumptions **(A1)**-**(A3)** and by the law of iterated expectations, the average structural function at time t among the treated $d = 1$ can be expressed as

$$ASF(\tilde{d}, z_1, d = 1) = \tag{6}$$

[32]Note that a standard exogeneity assumption as used for IV estimation of linear regressions such that the implied error term in equation **(A1)** is independent of the exogenous covariates can not hold because y_t is a discrete outcome variable, see Wooldridge (2014, p. 232).

[33]An alternative extension builds on the assumption that (u_t, ν) are jointly normally distributed. In this case, the vector of proposed control functions consists of three components: gr, gr^2, and the interaction between gr and linear predictions from the first stage probit estimation (Kimhi, 1999).

$$Pr_{b_t^d,u_t}\left(a_{t0}+z_1b_0+(b_{t0}+z_1(b_1-b_0))\tilde{d}+b_t^d\tilde{d}+u_t\geq 0 \mid d=1,z_1\right) .$$

where $Pr_{\xi}(\cdot)$ indicates a probability based on the distribution of ξ. Blundell and Powell (2003, 2004) and Wooldridge (2014) define the ASF for the entire sample. Because of our interest in the ATT, we restrict attention to the treated, and we define the ASF for the two potential outcomes conditional on $d=1$.

As suggested by Wooldridge (2014, section 6.4), we use the following flexible set of control functions $\hat{e}_2(d_i,z_i)=(\widehat{gr}_i,\widehat{gr}_i d_i,\widehat{gr}_i z_{i1})$, which enter as additional regressors in our fractional probit for employment (see equation 6). As motivated by Lee (1982), the interaction terms $\widehat{gr}_i z_{i1}$ account for deviations from the joint normality assumptions as imposed in Heckman (1978).[34] The interaction term $\widehat{gr}_i d_i$ accounts (as an approximation) for the random coefficient of the treatment dummy in the structural employment equation (5a), see e.g. Blundell et al. (2005, section 3.4.1) for the continuous outcome case.[35]

Under these assumptions, using the approach suggested in Wooldridge (2014, section 6.4), the average structural function for the treated can be expressed by integrating out the control functions e_2 as

$$ASF(\tilde{d},z_1,d=1)= \tag{7}$$

$$E_{e_2|d=1,z_1}\left\{Pr\left(a_{t0}+z_1b_0+(b_{t0}+z_1(b_1-b_0))\tilde{d}+b_t^d\tilde{d}+u_t>0 \mid d=1,z_1,e_2\right)\right\} ,$$

where $E_{\xi}[\cdot]$ indicates the expectation with respect to the distribution of ξ. Our estimation approach is based on the following insight: Equation (7) makes explicit that once the observed conditional expectation of y_t given (z,d,e_2) is estimated consistently, which in turn is implied by having sufficient variation in the instrumental variables z_2, identification of the average effect of treatment on the treated is feasible by integrating out the joint distribution of (z,e_2) among the treated.

For estimation purposes, we add estimated versions of the control functions e_2 to a second stage probit regression of employment, where we regress y_t on d, z_1, interactions

[34] Lee (1982) also suggested to add the squared generalized residuals $\widehat{gr}_i^2$ as further control variable to allow for a more flexible estimation approach. Doing so, we obtained rather implausible and noisy estimates. We do not report these results here because it turned the estimates are plagued by strong multicollinearity. We opted for keeping the interaction terms instead of adding $\widehat{gr}_i^2$. Detailed results are available upon request.

[35] Under joint normality of (ν,b_t^d,u_t), the coefficient of gr_i in the control function differs by treated status d_i because the linear projection of the joint error term $b_t^d d_i+u_t$ on ν differs by d_i. In the absence of the random coefficient part, i.e. $b_t=0$, and under joint normality of (ν,u_t), the generalized residual $\widehat{gr}_i$ has the same coefficient irrespective of the value d_i takes and thus $\widehat{gr}_i d_i$ should have a zero coefficient.

between d and z_1, as well as interactions between d and components of $\hat{e}_2$. The most general specification estimated is

$$\widehat{E}(y_{it} \mid d_i, z_{i1}, e_{i2}) = \tag{8}$$

$$\Phi\left(\hat{\delta}_{0t} + \hat{\delta}_{1t}d_i + z_1\hat{b}_0 + z_1\hat{\delta}_1 d_i + \hat{\omega}_0\widehat{gr}_i + \hat{\omega}_1\widehat{gr}_i d_i + z_{i1}\widehat{gr}_i\hat{\psi}\right) ,$$

where $\widehat{E}(y_{it}|.)$ is the estimated expected value, $\Phi(.)$ the standard normal distribution function, $\hat{\delta}_{0t}$, $\hat{\delta}_{1t}$ are time specific effects, and $\hat{b}_0$, $\hat{\delta}_1$, $\hat{\omega}_j$, $\hat{\psi}$ further coefficient estimates.

Based on the estimated equation (8), we estimate the ATT at time t by integrating out the distribution of $z_{i1}, \hat{e}_{i2}$ among the treated $d_i = 1$ as

$$\widehat{\tau}_{ATT,t} = \frac{1}{N_1}\sum_{d_i=1}\Big\{\Phi\left(\hat{\delta}_{0t} + \hat{\delta}_{1t} + z_1\hat{b}_0 + z_1\hat{\delta}_1 + \hat{\omega}_0\widehat{gr}_i + \hat{\omega}_1\widehat{gr}_i + z_{i1}\widehat{gr}_i\hat{\psi}\right) \tag{9}$$

$$- \Phi\left(\hat{\delta}_{0t} + z_1\hat{b}_0 + \hat{\omega}_0\widehat{gr}_i + z_{i1}\widehat{gr}_i\hat{\psi}\right)\Big\} .$$

Note that the same control function terms for treated individuals apply to the potential treatment and nontreatment state except for the coefficient $\hat{\omega}_1$ being 'switched off' in the nontreatment state (second term). This is because the $\hat{\omega}_1$-component of the individual selection effect does not have an impact in the nontreatment state (see Blundell et al. (2005), section 3.4.1).

Our empirical analysis pools the estimation of the time average of the outcome equation (5a) for year 1, years 2 to 3, and years 4 to 10. This results in the estimation of three separate fractional probit regressions. The dependent variables are the employment rates during year 1, years 2 to 3, and years 4 to 10, respectively.

Inference when estimating the fractional probit with control functions and the ATT is based on a clustered version of the weighted bootstrap (see Barbe and Bertail (1995)). The weighted bootstrap is based on resampling random weights with expected value one. This bootstrap approach eliminates potential estimation problems in the resamples, because in each resample a weighted fractional probit is implemented with strictly positive weights for all sample observations. The standard pairwise bootstrap may result in resamples with perfect collinearity of regressors or perfect predictions caused by the use of dummy variables. As suggested in Fitzenberger and Muehler (2015), we adapt the weighted bootstrap to estimate standard errors which are clustered at the district and year level. This is done by drawing the same weight within a cluster unit. Specifically, for the

weighted bootstrap, we draw the weights from a uniform distribution on the interval [0,2]. Therefore, the weights have a mean of one and a variance of 1/3. As drawing from this interval underestimates the variance by a factor of three, the obtained bootstrap variance-covariance matrix has to be multiplied by three.

A.3 List of variables

Variable	Description
	Instrument
Relative surplus	Difference between planned entries for the entire year and actual entries during the first six months per 1,000 eligible unemployed in July at the level of the local employment office (LEO)
	Individual characteristics
Female	Equal to one if female
Age	Age years dummies: **25–29**, 30–34, 35–39, 40–44, 45–50
Education	Education dummies: no vocational training degree, **vocational training degree**, uni/college degree, education unknown
Nationality	Equal to one if foreigner
Marital status	Equal to one if married
Children	Equal to one if at least one child in household
	Previous employment, former treatment, elapsed unemployment duration
Months employed	Dummies for being employed in month M (M=6, 12, 24) before current unemployment
Firm size	Dummies for $<$ **10 employees**, $\geq$ 10 and $<$20 employees, $\geq$ 20 and $<$ 50 employees, $\geq$ 50 and $<$ 200 employees, $\geq$ 200 and $<$ 500 employees, $\geq$ 500 employees/missing
Employment status	Dummies for apprentice, **blue collar worker**, white collar worker, worker at home/missing, part-time worker
Industry	Dummies for agriculture, basic materials, metal/vehicles/electronics, light industry, construction, **production oriented services/trade/banking**, consumer oriented, organizational and social services/missing
Occupation	Dummies for farmer and fisher, **manufacturing occupations**, technicians, service occupation, miners/others/missing
Wage	Log of daily earnings
Former participant	Dummies for participation in any active labor market training program reported in our data in year(s) Y (Y=1,2) before current unemployment
Elapsed unemployment	Dummies for **1**–12 months of elapsed unemployment duration
	Region and time information at program start
Region	Regional employment office: dummies for **Schleswig-Holstein and Hamburg (SHH)**, Lower Saxony and Bremen (NB), North Rhine-Westphalia (NW), Hesse (HE), Rhineland-Palatinate and Saarland (RPS), Baden-Württemberg (BW), North Bavaria (NBY), South Bavaria (SBY), Berlin (BE)
Year	Calendar year dummies for the time period from 1983-**1993**
Region $\times$ Year	Interaction terms between region and year
Month	Calendar month dummies for the months August-**November**
Summer vacation	Share of summer holidays in year and region; interaction term between summer vacation and August dummy; interaction term between summer vacation and calendar year
	Local labor market characteristics
Local unemployment rate	Local unemployment rate (LUR) in each month from January to July in the respective year of program start; LUR in last year before program start

Note: Variables in bold are the omitted category in the empirical analysis.