The Effectiveness of Public-Sponsored Training Revisited: The Importance of Data and Methodological Choices

Martin Biewen, *University of Tübingen, IZA, DIW Berlin*

Bernd Fitzenberger, *University of Freiburg, ZEW, IZA, ROA, IFS*

Aderonke Osikominu, *University of Hohenheim, CESifo, IZA*

Marie Paul, *University of Duisburg-Essen, RGS Econ*

This article revisits the effectiveness of public-sponsored training programs for Germany accounting for dynamic selection into heterogeneous programs. We carefully assess to what extent various aspects of our empirical strategy, such as conditioning flexibly on employment and benefit histories, the availability of rich data, handling of later program participations, and further methodological choices affect our estimates. Our results imply pronounced negative lock-in effects in the short run and positive medium-run effects on employment and earnings when job-seekers enroll after having been unemployed for some time. We find that data and specification issues can have a large effect.

We are indebted to Jeffrey Smith for many useful comments and suggestions. We also thank Peter Mueser and participants in numerous seminars for helpful comments. Special thanks go to Stefan Bender for his efforts to make available the

[*Journal of Labor Economics*, 2014, vol. 32, no. 4]

0734-306X/2014/3204-0007$10.00

I. Introduction

There has been an enormous interest in the evaluation of active labor market policies in general and of public-sponsored training in particular, both in the United States and Europe.[1] While earlier studies typically focused on the evaluation of a single program, recent developments in methodology and data access allow researchers to study the heterogeneity of treatment effects across different subgroups of people and the comparative effects of narrowly defined subprograms.[2] Detailed information on employment and earnings histories prior to program participation seems important to justify matching estimators for treatment effects that rely on a selection on observables assumption.[3] Accurate longitudinal information on labor market transitions is also useful to account for the dynamics of program assignment and to carefully align treated and comparison units in their elapsed unemployment experience.[4]

While methodological progress is likely to improve the policy relevance of scientific evaluations, it dramatically increases the range of possible choices researchers can make. A better understanding of how training programs work and to what extent evaluation results hinge on particular data features and methodological choices is crucial to assess research findings based on different empirical strategies and data sets. This article makes two contributions. On the substantive side, we revisit the effectiveness of two widely used

data used in this study. The study is part of the project Employment Effects of Further Training Programs, 2000–2002, an evaluation based on register data provided by the Institute for Employment Research, IAB (Die Beschäftigungswirkung der FbW-Maßnahmen 2000–2002 auf individueller Ebene—Eine Evaluation auf Basis der prozessproduzierten Daten des IAB; IAB project number 6–531.1A, joint project with University of St. Gallen and the IAB). We gratefully acknowledge financial and material support by the IAB. Bernd Fitzenberger gratefully acknowledges financial support by the German Research Foundation (DFG) through the research project Statistical Modelling of Labor Market Processes in Misclassified Administrative Labor Market Data. Aderonke Osikominu gratefully acknowledges financial support from the German Excellence Initiative and the Austrian Science Fund (FWF grant no. S 10304-G16). The usual caveat applies. Throughout the article, reference is made to material in a supplemental online appendix. This can be found by using a link to a PDF that is located in the online version of the article. Contact the corresponding author, Bernd Fitzenberger, at bernd.fitzenberger@vwl.uni-freiburg.de.

[1] For comprehensive overviews, see Heckman, LaLonde, and Smith (1999), Martin (2000), Martin and Grubb (2001), Kluve and Schmidt (2002), Carcillo and Grubb (2006), and Card, Kluve, and Weber (2010).

[2] Examples of evaluations involving multiple comparisons of different programs are Gerfin and Lechner (2002), Lechner (2002), Larsson (2003), Hardoy (2005), Dorsett (2006), Dyke et al. (2006), Frölich (2008), and Sianesi (2008).

[3] See, e.g., Heckman et al. (1998, 1999), Heckman and Smith (1999), and Dolton and Smith (2011).

[4] See, e.g., Sianesi (2004), Fredriksson and Johansson (2008), Crépon et al. (2009), Dolton and Smith (2011), and Osikominu (2013).

public-sponsored training programs—the first one involving *comprehensive classroom further training* and the second one focusing on *short-term activation and reemployment*—using a dynamic propensity score matching approach. On the methodological side, we undertake a detailed sensitivity analysis to investigate the importance of data features and specification choices, as well as econometric modeling strategies, in shaping estimation results.

Our first methodological contribution relates to the availability of rich data. We have access to unique data for Germany that merges information from different administrative registers. They contain precise and extensive information on individual employment and benefit histories covering more than 10 years before and about 2.5 years after treatment start, detailed information on participation in all active labor market programs, and detailed profiles of job-seekers (i.e., personal characteristics and goals of job search) that are reported by the caseworkers at the local employment agencies. Using different balancing tests, we first show that the information on personal characteristics and individual employment and benefit histories in our data balances individual characteristics and pretreatment outcomes in our treatment and comparison groups well. We then carry out a careful sensitivity analysis that compares our benchmark evaluation results to different variations in which we omit important aspects of our data or our specifications. We believe that such an analysis provides important information for the evaluation of labor market programs in other countries. For example, many data sets used for evaluation purposes lack the information about what other programs comparison group members might participate in, for example, that of Mueser, Troske, and Gorislavsky (2007). We investigate the sensitivity of the results if we proceed as if we did not have this information. Similarly, in the spirit of Card and Sullivan (1988), Heckman and Smith (1999), and Dolton and Smith (2011), we examine what difference conditioning on various aspects of the employment and benefit history makes. In order to assess to what extent evaluation results hinge on the availability of rich personal information, we compare our results to several specifications in which we omit information, for example, about the motivation of participants or on health and other characteristics that are often not available in data sets used for evaluation purposes.[5]

[5] Complementary to our first methodological contribution, Lechner and Wunsch (2011) examine the importance of different sets of conditioning variables using the same administrative data and propensity score matching in combination with matching on actual and hypothetical program starts. Their study involves an Empirical Monte Carlo analysis that presumes a known data-generating process using real data and the validity of the conditional independence assumption. Our analysis, in contrast, investigates the sensitivity of the estimated treatment effects. In another Empirical Monte Carlo analysis, Huber, Lechner, and Wunsch (2010) investigate the performance of different propensity score matching estimators.

Our second methodological contribution relates to the implementation of a matching approach under dynamic treatment assignment. People who become unemployed may at some later point take up a job or participate in one of the training programs under investigation or in some other labor market program. According to many labor market policy regulations, as is the case also in Germany, these three options are to be viewed as competing risks. In particular, somebody who has found employment again is not eligible for program participation anymore. In this setting, a simple static evaluation approach in which the unemployed receiving treatment are compared to the unemployed who never receive treatment is not feasible. Defining nonparticipants as never participants implies conditioning on future outcomes because for these persons the exit to employment has occurred before the potential start of treatment (Fredriksson and Johansson 2008). Put differently, excluding comparison individuals who receive treatment in the future excludes possible sequences of events that may have occurred under the counterfactual situation that a treated individual had not started the program at the particular point of time observed. Similarly, we may only observe a treatment starting at some point of time during an unemployment spell because the unemployed did not happen to find a job before. Following Sianesi (2004), we define treatment parameters conditional on the unemployment experience at the time of treatment start and include all those in the comparison group for a particular treatment start who do not enroll at the elapsed unemployment experience under consideration. As a consequence, the future outcomes of comparison units may be affected by the effect of the treatment. In our sensitivity analysis, we investigate the sensitivity of the estimated treatment effects to not using the information on future program participation and on participation in other labor market programs, as well as to not aligning treated and comparison persons by elapsed unemployment duration. We further compare the results obtained in our benchmark approach to an alternative approach suggested by Lechner (1999). The latter assigns to each nontreated individual who does not receive the treatment of interest within a fixed time window (e.g., 12 months) a hypothetical program start date that is simulated based on covariates observed at the start of the unemployment spell. A nontreated individual is then only used as comparison if he or she is still unemployed before the hypothetical starting date. In an attempt to align treated and comparison units in their prior unemployment experience, similarity of the actual and hypothetical program start dates is introduced as a further matching requirement. Our second methodological sensitivity analysis thus provides new insights on the importance of econometric modeling choices to control for dynamic selection effects.

Finally, as our substantive contribution, we assess the effectiveness of the two training programs during a time when resources were shifted away from longer-term training programs to short-term programs and to more

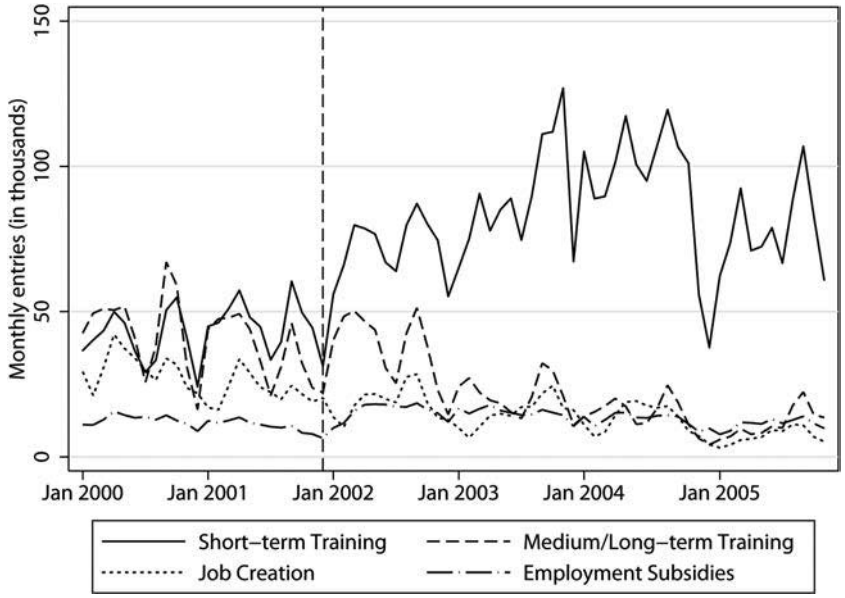

FIG. 1.—Active labor market programs in Germany. Source: Bundesagentur für Arbeit (2002b, 2005b), own calculations. The category "medium/long-term training" represents the programs we also call "further training" and "retraining" in the text.

job-oriented programs (see Bundesagentur für Arbeit [2005c], and fig. 1). Given the wide range of training programs implemented in Germany, our comparison of different forms of training may also hold lessons for other countries that employ similar programs. Short-term training is particularly interesting because, in addition to the provision of skills, it includes elements of job search assistance, as well as profiling and monitoring of the unemployed (see the literature on such programs reviewed in the next section). Our results show sizable lock-in effects for all programs in the short run. In the medium run, we find positive effects on employment and earnings when job-seekers enroll after having been unemployed for some time. Although both types of training programs generate significant earnings gains in the cases in which there are also positive employment effects, earnings gains induced by classroom further training attain €200 per month and are generally larger than those of short-term training (€70 per month for men and €175 for women). We find little evidence for additional impact heterogeneity after stratifying by gender and elapsed unemployment duration.

The remainder of this article proceeds as follows. Section II reviews the training literature. Section III describes the institutional background and data. We introduce our econometric approach in Section IV. The evalu-

ation results and our sensitivity analysis are presented in Sections V and VII. Section VII concludes. An appendix, available online in PDF form, provides complementary empirical results and background information.

II. Review of the Training Literature

The earlier literature for the United States (see the reviews by Barnow [1987], Bloom et al. [1997], and Heckman et al. [1999]) distinguishes classroom training, on-the-job training, and work experience. Programs tend to be more effective for women than for men, and on-the-job training tends to be more effective than classroom training and work experience. More recent evaluations for the United States make use of advances in econometric methods, often investigate impacts over a longer time period, and put a stronger emphasis on the heterogeneous effects of different training programs (Dyke et al. 2006; Hotz, Imbens, and Klerman 2006; Heinrich et al. 2010). These studies investigate programs involving job search assistance, improvement in job readiness, or training, and they compare longer to shorter programs. Dyke et al. (2006) and Hotz et al. (2006) analyze training programs for a specific group of welfare recipients, whereas our study concerns training among a sample of prime-age unemployed who were employed in the recent past. Heinrich et al. (2010) find positive medium-run to long-run effects for the broader group of participants in programs under the Workforce Investment Act. Their findings imply higher gains for participants in general programs for adults than for participants from the group of displaced workers.

In a study for Canada, Park et al. (1996) distinguish training programs for the long-term unemployed, reentry programs for women, and programs addressing specific skill shortages. The study finds that only the latter two involve positive earnings effects. Hui and Smith (2003a, 2003b) distinguish training programs by the source of financing (private, public, employer). The study concludes that the quality of their data is not sufficient to provide credible estimates. Human Resources and Skills Development Canada (2004) provides a comprehensive evaluation of employment benefits and support measures within the joint Labor Market Development Agreement (LMDA) in British Columbia. The measures evaluated include a training subsidy called skills development employment benefits. The results suggest that all measures including the skills development employment benefits have positive earnings and employment impacts for active claimants of unemployment benefits but generally not for former claimants. For the United Kingdom, Dorsett (2006) analyzes the relative effectiveness of the five options offered under the New Deal for Young People. The five options are remaining on the gateway of intensive job search, subsidized full-time employment, full-time education or training, work placement in a voluntary sector, and placement in the so-called Environmental Task Force. In most

cases, entering one of the other options was not found to be more effective than intensive job search.

As one of the first studies providing a pairwise evaluation of multiple, mutually exclusive treatments (Imbens 2000; Lechner 2001), Gerfin and Lechner (2002) analyze the comparative effectiveness of five types of public-sponsored training programs in Switzerland, with durations ranging between 5 and 13 weeks, using high-quality administrative data. The study finds that, 1 year after program start, the employment rate of participants is lower than that of comparable nonparticipants. Longer, more intensive training programs show less negative effects than shorter ones.

For the Nordic countries, a number of studies investigate the effects of public-sponsored training programs, often based on high-quality administrative data. Larsson (2003) finds zero or even negative effects of two youth labor market programs in Sweden, a subsidized work program and a training program, compared to no treatment. Sianesi (2008) analyzes the comparative effectiveness of different labor market programs in Sweden, among them vocational classroom training, work practice (subsidized employment including elements of training), and trainee replacement (training on-the-job mostly in the public sector). She finds moderate positive effects for trainee placement but no positive effects for other forms of training. For Norway, Hardoy (2005) and Jespersen, Munch, and Skipper (2008) find negative effects of classroom training. The evidence on more practically oriented training is mixed.

A number of studies use high-quality administrative data for Germany during the 1990s with an evaluation period after program start of up to 8 years. Lechner, Miquel, and Wunsch (2007, 2011) distinguish between medium-term programs (mean duration 4 months), longer programs (mean duration 9–12 months), and long programs with specific contents. Most of the programs show positive effects in the long run, even in East Germany. An important finding is that medium-term programs outperform longer programs, as they exhibit a much shorter lock-in period with otherwise similar employment effects after the end of the program. Similar findings are obtained by Fitzenberger and Speckesser (2007), Fitzenberger and Völter (2007), and Fitzenberger, Osikominu, and Völter (2008) based on the same data but using a different methodological approach building on Sianesi (2004).

The administrative data for Germany during the 1990s are yet still much less informative than the administrative data used in this study for the 2000s. For instance, the data for the 1990s only contain information on periods with unemployment benefits but not about being registered as unemployed. Furthermore, the older data involve less rich information on personal characteristics and no information about other active labor market programs apart from training. As to be expected from the evidence reported in Sianesi (2008) for Sweden, a country with a comprehensive set

of active labor market policies like Germany, information about other programs is important, as our analysis will show. Our data set is not unmatched in terms of informational content compared to data sets used for the evaluation of public-sponsored training programs in other countries. Our impression is that there are administrative data sets available for the Scandinavian countries, Switzerland, and Austria that are more or less just as rich in informational content as the recent administrative data for Germany. In contrast, the data sets used for the United States, Canada, the United Kingdom, and most other countries seem typically much less informative in comparison.

A further drawback of the studies for Germany for the 1990s is that they do not analyze short-term training programs, which were abolished in 1993.[6] Since their reintroduction in 1998, short-term training programs (Trainingsmaßnahmen) have by now become the most important active labor market program regarding the number of participants (see Sec. III.A). Evaluations for other countries (Blundell et al. 2004; Weber and Hofer 2004; Crépon, Dejemeppe, and Gurgand 2005; Fougère, Pradel, and Roger 2005; van den Berg and van der Klaauw 2006) and policy discussions (Martin and Grubb 2001; OECD 2005) suggest that short-term training programs are superior to longer training programs because the former activate the unemployed without a long lock-in effect. Job search assistance may be an inexpensive way to increase the employment chances of job-seekers. Using data for the 2000s, Hujer, Thomsen, and Zeiss (2006) find that participation in short-term training reduces the unemployment duration of West German job-seekers.

The studies mentioned so far for Germany do not compare short-term training to other programs. Using the data for the 2000s, Schneider et al. (2006) find that shorter training programs, which are longer than short-term training programs, tend to be more effective than longer ones. Osikominu (2013) uses the same administrative data as the current article but a duration framework in continuous time. They find that short-term training reduces the remaining time in unemployment and that it has moderate positive effects on subsequent job stability. Longer training programs initially prolong the remaining time in unemployment but former participants have substantially more stable employment spells and earn more. Using the same administrative data as we do, Wunsch and Lechner (2008) and Lechner and Wunsch (2009) estimate differential effects of various programs including short-term training.[7] For West Germany, partly in contrast to our results in this study, Wunsch and Lechner (2008) find negative lock-in effects for all

[6] An exception is Fitzenberger et al. (2013), which compares the effects of different types of short-term training for the 1980s, the early 1990s, and the early 2000s and which finds modest positive employment effects in some but not all cases.

[7] Similar to Lechner and Wunsch (2009), Biewen et al. (2007) do not find significantly positive employment effects for East Germany.

programs in the short run and no—or even negative—medium-run effects on employment and earnings. We take the differences to our study as a motivation to analyze the sensitivity of results with regard to methodological choices. We note here three important methodological aspects of the Wunsch and Lechner study.[8] First, the comparison group is defined based on nonparticipation during an observation window of 18 months (in an additional sensitivity analysis, a time window of 12 months is used). Second, Wunsch and Lechner (2008) simulate hypothetical program start dates for controls and use these dates for the alignment of treated and controls in time. This presumes matching on the duration until treatment, which is unknown for the controls. Third, in addition to the start dates of the program, Wunsch and Lechner (2008) match on an estimated propensity score that does not change over the course of the unemployment spell.

III. Background and Data

A. Training as Part of Active Labor Market Policy

The main goal of German active labor market policy is to permanently reintegrate unemployed individuals (and individuals who are at risk of becoming unemployed) back into employment. There exists a wide range of different programs, such as wage subsidies, job creation schemes, youth programs, programs to promote self-employment, and training programs (see fig. 1 for an overview). Training programs have traditionally been the most important part of active labor market policy. This article focuses on training programs during the time period 2000–2002.[9] There are three main types of training: *short-term training* (Trainingsmaßnahmen), *further training* (Berufliche Weiterbildung), and *retraining* (Umschulung).[10] Apart from the fact that all three types of training require full-time participation, they differ considerably in length and content.

Short-term training programs last only 2–12 weeks (the mean duration is slightly over 4 weeks; see table 1), and they typically have one or several of the following three goals. A first goal is to assess job-seekers' labor market opportunities and their suitability for different jobs. This may also entail profiling and developing a strategy to find a job. A second goal is to test job-seekers' willingness to work and to improve their job search skills.

[8] Another difference between the evaluation sample in Wunsch and Lechner (2008) and ours is that it does not include programs financed under the heading "Freie Förderung" and by the ESF (European Social Fund). Further, it restricts the analysis to workers who receive unemployment assistance or unemployment benefits.

[9] We focus on the period before the Hartz-Reforms were implemented that changed some of the rules on training programs (Schneider et al. 2006).

[10] In addition, there are specific training programs for disadvantaged youths and disabled persons as well as German language courses.

Table 1
Average Expenditures on Training and Unemployment Compensation and Average Training Durations in Germany

	2000	2001	2002	2003	2004
Short-term training:					
Enrollment length	1.2	1.1	.9	1.0	.9
Training costs	580	570	658	538	421
Further training and retraining:					
Enrollment length	8.2	9.3	9.1	10.5	10.7
Training costs	664	664	681	631	627
Subsistence allowance	1,152	1,178	1,188	1,156	1,150
Unemployment compensation:					
Unemployment benefits	1,160	1,189	1,185	1,261	1,313
Unemployment assistance	753	721	727	691	713

SOURCE.—Bundesagentur für Arbeit (2001, 2002a, 2005a), own calculations.
NOTE.—Rows labeled "Enrollment length" display the average duration of the program in months. The rows labeled "Training costs" and "Subsistence allowance" and the two rows under the heading "Unemployment Compensation" contain the average monthly expenditures (in euros) divided by the average monthly number of participants/claimants. Expenditures on subsistence allowance, unemployment benefits, and unemployment assistance include social security contributions.

This may involve job application training. The third goal is the provision of specific skills that are necessary to improve job-seeker's labor market prospects, for example, through computer courses or commercial training. In 2001, about a fifth of short-term training programs focused on the first and second goals, respectively, and 28% on the third; 31% served more than one of these goals (Kurtz 2003).

The more substantial *further training programs* typically last between several months and 1 year, thus representing medium-term programs. Their goals are to maintain, update, adjust, and extend occupational skills. The programs cover a wide range of fields and may also comprise practical elements such as on-the-job training or working in practice firms. Typical examples include training on marketing and sales strategies, computer-assisted bookkeeping, and operating construction machines, and completing specialist courses in specific legal fields. Depending on their practical content, we distinguish between *classroom further training* and *practical further training*. Finally, *retraining* programs involve training on a new vocational degree according to the German system of vocational education. They last 2–3 years.

Eligibility for one of the training programs requires registration as a job-seeker at the local labor office. This involves a counseling interview with the caseworker. For participation in the longer further training and retraining programs, individuals also have to fulfill a minimum work requirement of 1 year, and they must be entitled to unemployment benefits. However, there are a number of exceptions. The binding criterion is that the training program has to be considered necessary in order for the job-seeker to take up a job, for example, because the training is required to meet the hiring standards of the job. Training programs are usually assigned by

the caseworker depending on the regional supply of upcoming training slots. A participation in training may take place at any point in time during the unemployment spell. Job-seekers have no entitlements regarding participation. A program assignment is compulsory for the job-seeker, and noncompliance may entail benefit sanctions and the exclusion from further services. The employment agency covers all direct training costs. In addition, participants in short-term training may continue to receive unemployment benefits or means-tested unemployment assistance, if eligible. Participants in further training and retraining usually receive a subsistence allowance of the same amount as unemployment benefits or unemployment assistance, provided they fulfill the minimum work requirement.

Table 1 shows that the average monthly training costs per participant are lower for short-term training courses (€570 in 2001) than for the medium- and long-term programs (€664). Given that the average length of short-term training is only 1.1 months while that of the longer programs is 9.3 months, this results in 10 times higher costs for the medium- and long-term programs as opposed to short-term training (€6,175 vs. €627). In light of their substantially lower costs, participation in short-term training programs has been expanded since 2002 compared to longer-term programs (fig. 1). Short-term training has become the largest training program regarding the number of participants. These stark differences in costs and durations motivate our comparison of the two types of training programs.

B. Administrative Database Used

Our study uses a new and exceptionally rich administrative database, the Integrated Employment Biographies Sample (IEBS), provided by the Research Data Center of the Federal Employment Agency.[11] The IEBS is a 2.2% random sample from a data file combining individual records out of four different administrative registers:. (i) the Employment History (Beschäftigten-Historik), (ii) the Benefit Recipient History (Leistungsempfänger-Historik), (iii) the Database of Registered Job-Seekers (Bewerberangebot), and (iv) the Database of Program Participants (Maßnahme-Teilnehmer-Gesamtdatenbank). Start and end dates of the different labor market episodes are measured with daily precision. In addition to individual labor market histories, the IEBS contains extensive information on personal characteristics, occupation characteristics, and job characteristics, as well as regional identifiers, which allows us to merge, for example, the unemployment rate at the county level.[12] For

[11] See Jacobebbinghaus and Seth (2007) and http://fdz.iab.de/en.aspx. Being among the first to use the IEBS, we performed extensive consistency checks; see Bender et al. (2004, 2005), Fitzenberger, Osikominu, and Völter (2006), and Waller (2008).

[12] Counties do not necessarily coincide with local labor markets. Nevertheless, the county-level unemployment rate is a good proxy for the local labor market conditions.

evaluation purposes, a rich set of covariates is essential to reconstruct the circumstances affecting program participation and labor market outcomes.

The Employment History involves register data for the time period January 1990 to December 2004 on employment subject to social security contributions.[13] Earnings before taxes are collected on an annual basis in a given employment relationship if no changes in relevant information, such as a change of the health insurance, occur within the year. Specifically, for somebody continuously working with a given employer, we observe a spell that lasts from January 1 to December 31 of each year, together with total earnings during each spell. This allows us to calculate earnings per calendar day as well as over more aggregated time intervals. We use these data to construct variables that reflect the employment status and wages during the pretreatment and posttreatment periods, spanning about 10 years in total. We only consider unsubsidized employment spells for our outcome measures. In addition, we use information on year of birth, gender, occupation, and industry.

The Benefit Recipient History includes spells of all unemployment benefit, unemployment assistance, and subsistence allowance payments between January 1990 and June 2005, together with some personal characteristics. It provides information on periods during which people are not employed and therefore not covered by the Employment History. We use the information on benefit payments to construct individual benefit histories dating back for 7 years. Moreover, we use information on reasons for suspensions of payments to proxy a lack of motivation to find a job.

The Database of Registered Job-Seekers contains information on job search during the time period January 1997 to June 2005.[14] First, the data source provides additional information about the labor market status of a person, that is, whether somebody is looking for a job while employed or nonemployed or whether somebody is temporarily sick while being registered as unemployed. Second, it includes rich information about personal characteristics, for example, educational attainment, nationality, family situation, and health status, as well as characteristics of the last job and the goals of the job search.

The Database of Program Participants contains detailed information on participation in all public-sponsored labor market programs (5-digit program codes), as well as some additional variables such as the planned end date. This register covers the time period January 2000 to July 2005. We use the information on participation in job creation schemes and wage

[13] This is equivalent to about 80% of the employed persons (Bender, Haas, and Klose 2000). The main groups not covered are the self-employed and civil servants.

[14] According to the data manual for the IEBS, the Database of Registered Job-Seekers is complete only from 2000 onward. However, this does not seem to be an issue for our analysis as the share of missings in our retrospective variables does not vary systematically over time.

subsidy programs to identify subsidized employment spells in the Employment History.

C. Evaluation Sample and Training Programs Analyzed

Our evaluation approach involves comparisons between a group of people receiving a program of interest and a group of people who do not receive the program but who could do so in principle. Correspondingly, one has to determine first who is a potential program participant. We focus on individuals who become unemployed after having been continuously employed for a while, instead of a stock sample of observed unemployed at a given point of time. This way we exclude those individuals who have been out of the labor force and who register as unemployed just because they want to participate in a training program. In interviews, caseworkers told us that especially women who are returning from maternity leave, or who are divorced, or who are university graduates, and who are having difficulties finding a job, contact the local employment office inquiring about the possibility to participate in programs. Thus, an evaluation sample based on the observed unemployment status at a particular point in time (instead of an inflow sample from employment into unemployment) has important disadvantages. Since participants returning from out of the labor force are not registered as unemployed before the start of training, the corresponding comparison persons, who are still out of the labor force, are not recorded in the data. Even if it were possible to observe these people in the data, this would entail the danger that these individuals are not considering taking up a job, which is the main purpose of active labor market programs. Therefore, we focus on individuals who are attached to the labor market because they were employed in the recent past. This enables us to construct a suitable comparison group based on the information in the data. Furthermore, the beginning of the unemployment spell defines a natural time scale to align treated and nontreated individuals.

Our sample of inflows into unemployment comprises West Germans who become unemployed between the beginning of February 2000 and the end of January 2002, after having been continuously employed for at least 3 months. Entering unemployment is defined as ending regular, nonsubsidized, full-time or part-time employment and subsequently being in contact with the employment office (not necessarily immediately), as reflected by a job search spell, benefit receipt, or program participation.[15] In order to exclude individuals eligible for specific youth programs or for early retirement schemes, we only consider persons who are aged between 25 and 53 years at the beginning of their unemployment spell. Our evaluation fo-

[15] About 10% of the individuals in our inflow sample appear with more than one unemployment spell according to our inflow definition. We take account of this when calculating standard errors (see Sec. IV).

cuses on the first training program that is attended in the course of a given unemployment spell.

We report results for two different types of training that largely follow the legal grouping of program types: *short-term training* (STT) and *classroom further training* (CFT; see Sec. III.A).[16] In a few cases, we also included programs with similar contents and planned durations but with a somewhat different legal basis into our program types.[17] While we display treatment effects only for STT and CFT, we consider participation in other programs, for example, practical further training and subsidized employment, when defining the comparison groups (see table 2). That is, participants in other programs are not counted as in open unemployment. We estimate program effects separately for men and women and for three strata representing different durations of elapsed unemployment. This results in a total of six evaluation samples (stratum 1/stratum 2/stratum 3, male/female). Table 1 gives an overview of the sample sizes and the transitions from open unemployment. Figure 2 provides descriptive information on the duration of different program types. The STT programs last on average about 1 month, while the CFT programs last on average 7.5 months, with spikes in the distribution at 6 and 12 months.

IV. Econometric Approach

A. Evaluating Multiple Treatments in a Dynamic Context

Our empirical analysis is based upon the potential-outcome approach to causality (see Splawa-Neyman 1923/1990; Roy 1951; Rubin 1974; also see the survey of Heckman et al. 1999). Imbens (2000) and Lechner (2001) extend this framework to allow for multiple, exclusive treatments. Our analysis distinguishes two different treatments, STT and CFT, as well as a nontreatment state. Let Y^k represent the potential outcome associated with training program $k = 1, 2$ and Y^0 represent the potential outcome when not participating in any program. For each individual, only one of the three potential outcomes is observed. We focus on the first treatment in a given unemployment spell. Our goal is to estimate the average treatment effect on the treated (ATT) of receiving treatment, $k = 1, 2$, as the first treatment both against nonparticipation $l = 0$ ("treatment" vs. "searching") and against treatment l, with $k \neq l$ and $k, l \neq 0$ (differential effect of two programs).

In our context, participation in training is possible at any point in time during the unemployment spell. A job-seeker who has not yet enrolled at a given point in time may join at some later point in time provided he

[16] In Biewen et al. (2007), we also report results for *practical further training* and *retraining* as well as for East Germany.

[17] We assigned some programs in the legal category of *Discretionary Support* (Freie Förderung) and programs financed by the *European Social Fund* (Europäischer Sozialfonds) into one of our program categories.

Table 2
Sample Sizes, Program Participation, and Transitions by Stratum

	Stratum 1	Stratum 2	Stratum 3
Males:			
Unemployed at start of stratum	32,172	20,335	13,241
Short-term training	912	547	662
Classroom further training	389	251	270
Practical further training	86	71	107
Retraining	263	89	99
Other program	1,171	648	755
No program participation	29,351	18,729	11,348
Exit to employment	9,016	5,488	3,113
Unemployed until end of stratum	20,335	13,241	8,235
Females:			
Unemployed at start of stratum	20,746	13,800	10,096
Short-term training	693	409	497
Classroom further training	344	194	201
Practical further training	88	61	85
Retraining	262	55	72
Other program	671	332	446
No program participation	18,409	12,749	8,795
Exit to employment	4,609	2,653	1,939
Unemployed until end of stratum	13,800	10,096	6,856

NOTE.—Stratum 1 refers to months 1–3 of unemployment, stratum 2 to months 4–6, and stratum 3 to months 7–12, respectively. The first row shows the number of people who are still unemployed at the beginning of the stratum and who have not received a program before. The subsequent rows show their transitions into programs and employment and the survivors in unemployment until the end of the stratum.

or she remains unemployed up to that time. In such a setting, persons with longer completed unemployment durations are more likely to end up receiving treatment than persons with shorter completed unemployment durations. Fredriksson and Johansson (2008) show that in this case a static evaluation analysis that assumes that the treatment is administered only once yields biased treatment effects because the definition of the comparison group conditions on future outcomes.[18] We follow the approach suggested in Sianesi (2004, 2008) to extend the static multiple treatment approach to a dynamic setting. Sianesi focuses on the effect of participating after a given unemployment experience (i.e., period of eligibility) versus not participating up to that point in time and continuing to search for a job. Thus, treatment effects are defined conditional on a given starting date during the unemployment spell, and the treated and comparison individuals are aligned by the elapsed unemployment duration. Note that individuals in the nontreatment group may enroll in training at some later point during the unemployment spell. This treatment parameter (treatment vs. searching)

[18] While Fredriksson and Johansson (2008) then estimate the effect on the hazard of leaving unemployment, our outcome variables of interest are the monthly employment status and earnings.

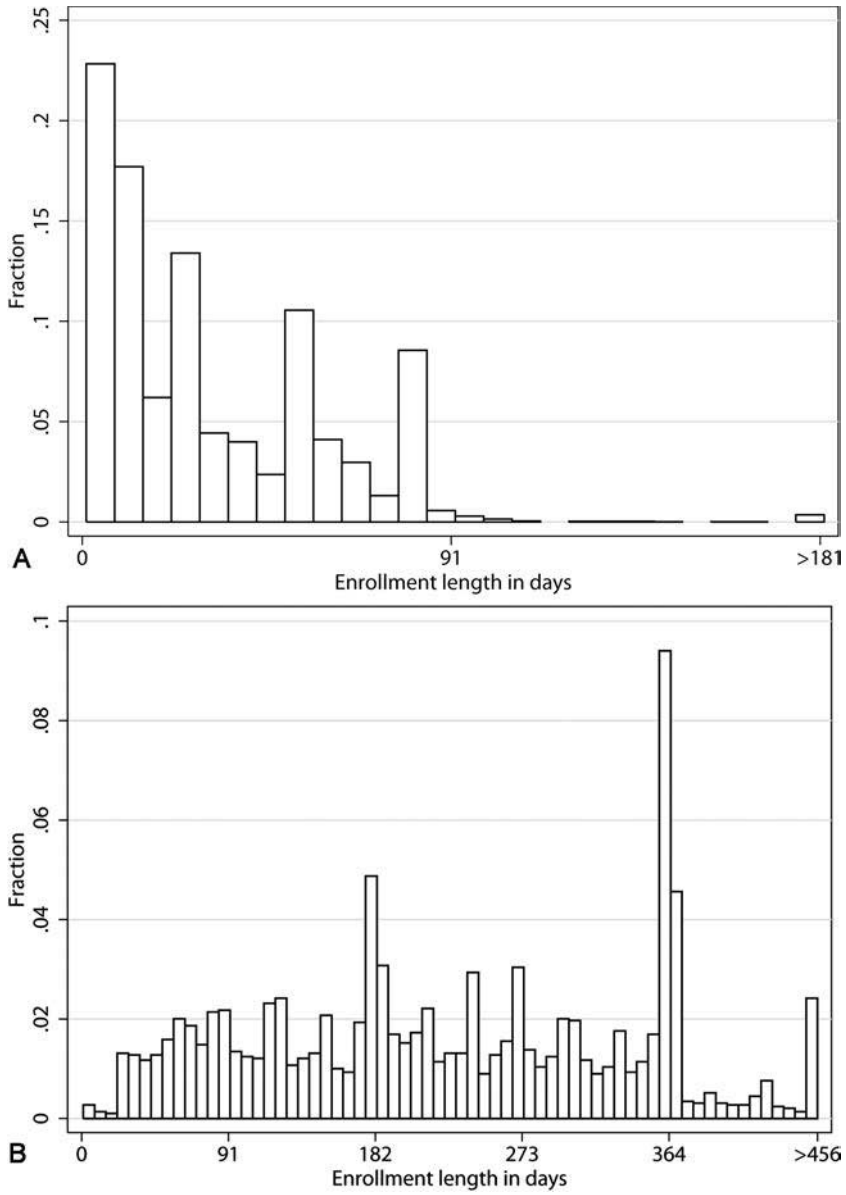

FIG. 2.—*A*, Short-term training; *B*, Classroom further training. Distribution of program durations in the evaluation sample. The mean duration of short-term training is 32 days, the median 26. The mean duration of classroom further training is 226 days, the median 217.

mirrors the decision problem of the caseworker and the unemployed who recurrently during the unemployment spell decide whether to start any of the training programs now or to postpone possible participation to the future in case job search will not be successful. When discussing the policy implications of our estimates, one has to be careful because those nontreated individuals in one stratum who receive treatment later experience the impact of the treatment.

We distinguish between treatment starting during months 1–3 of the unemployment spell (stratum 1), treatment starting during months 4–6 (stratum 2), and treatment starting during months 7–12 (stratum 3). We experimented with shorter strata but found that the number of treated and comparison individuals became too small (thus resulting in higher variance of the estimates) to implement our matching approach that relies on both exact matching and kernel matching (see below). Also due to concerns regarding small sample sizes, we do not evaluate training programs starting more than 12 months after the beginning of unemployment.

We evaluate treatments conditional on the unemployment spell lasting at least until the start of treatment k. Let $u = 1, 2, \ldots$ denote the month in which treatment k starts in the unemployment spell. For instance $u = 2$ means that treatment starts in month 2 of unemployment; $\tau = 0, 1, 2, \ldots$, counts the months since the beginning of treatment k; and $Y^k(u, \tau)$ is then the potential outcome in period $u + \tau$ for treatment k starting in month u. Similarly, $Y^l(\tilde{u},\ \tau - (\tilde{u} - u))$ is the alternative potential outcome in period $\tilde{u} + [\tau - (\tilde{u} - u)]$ associated with the alternative treatment l starting in period $\tilde{u}$. The ATT parameter for treatment k against the alternative l ($k \in \{1, 2\}$, $l \in \{0, 1, 2\}$, $k \neq l$) is given by

$$\begin{aligned}\theta(k, l; u, \tau) = {} & \mathrm{E}(Y^k(u, \tau) | U^o = u - 1, T_u = k) \\ & - \mathrm{E}_{u \le \tilde{u} \le \bar{u}_s}(Y^l(\tilde{u}, \tau - (\tilde{u} - u)) | U^o = \tilde{u} \ge u - 1,\ T_u = k),\end{aligned} \tag{1}$$

where U^o denotes the random time spent in open (i.e., untreated) unemployment and $T_u \in \{0, 1, 2\}$ the treatment status in period u; $\underline{u}_s$ ($\bar{u}_s$) denotes the start (end) of stratum s, $s \in \{1, 2, 3\}$; and $\mathrm{E}_{u \le \tilde{u} \le \bar{u}_s}(\cdot)$ the expectation with respect to the distribution of $\tilde{u}$ in the remainder of the stratum. Thus, for treatment k starting in period u, we require that possible comparison individuals receiving treatment l have spent the same amount of time in open unemployment as of period $u - 1$ and receive treatment l no earlier during the stratum considered than the treated individual. For $l = 0$, we have $\tilde{u} = u$ and the comparison group comprises all individuals whose unemployment spell lasts at least $u - 1$ periods and who do not participate in any program during the stratum considered. This entails the possibility of exit to employment before the end of the stratum. For $l \ge 1$, $\tilde{u}$ can take any value in the interval $[u, \bar{u}_s]$ and $\tau - (\tilde{u} - u)$ counts the months since start of treatment l,

thus aligning elapsed unemployment duration. In this way, we compare participation in one program to possibly delayed participation in another program later during the stratum.

We display treatment effects by stratum only, averaging the period-specific treatment effects, $\theta(k, l; u, \tau)$, with respect to the distribution of starting dates in the stratum:

$$\theta(k, l; s, \tau) = \sum_{u} f_s(u)\theta(k, l; u, \tau),$$

where $f_s(u)$, $u \in \{\underline{u}_s, \ldots, \bar{u}_s\}$, is the distribution of starting dates in stratum s.

We assume the following dynamic version of the conditional mean independence assumption (DCIA):

$$\begin{aligned} &\mathrm{E}_{u \le \tilde{u} \le \bar{u}_s}(Y^l(\tilde{u}, \tau - (\tilde{u} - u))|U^o = \tilde{u} \ge u - 1, T_u = k, \boldsymbol{X}_s) = \\ &\mathrm{E}_{u \le \tilde{u} \le \bar{u}_s}(Y^l(\tilde{u}, \tau - (\tilde{u} - u))|U^o = \tilde{u} \ge u - 1, T_{\tilde{u}} = l, \boldsymbol{X}_s), \end{aligned} \tag{2}$$

where $\boldsymbol{X}_s$ denotes a vector of covariates that are time invariant within a stratum. We effectively assume that, conditional on $\boldsymbol{X}_s$ and on remaining in open unemployment at least until period $u - 1$, individuals are comparable in their outcome for treatment l occurring in month $\tilde{u}$ between u and $\bar{u}_s$. As usual, implicit is the assumption that potential outcomes are independent across individuals, ruling out general equilibrium effects.

B. Combining Exact Matching and Kernel Matching

Building on Rosenbaum and Rubin's (1983) result on the balancing property of the propensity score in the case of a binary treatment, Lechner (2001) shows that the conditional probability of treatment k, given that the individual receives treatment k or treatment l, $P^{k|kl}(\boldsymbol{X}_s)$, exhibits an analogous balancing property for the comparison of program k versus l. This allows us to apply standard binary propensity score matching based on the sample of individuals participating in either program k or l. For this subsample, we simply estimate the probability of treatment k and then apply a bivariate extension of standard propensity score matching techniques. Implicitly, we assume that matching on the elapsed unemployment duration, prior employment history and calendar time at the start of the spell accounts for any remaining selection within the stratum.

To account for the dynamic nature of treatment assignment, we estimate the probability of receiving treatment k versus l in a given stratum using all the individuals in group k and l who are still unemployed at the beginning of the stratum considered. For treatment during months 1–3, we take the total sample of unemployed, who participate in k or l during months 1–3 (stratum 1), and estimate a Probit model for participation in k. For $l = 0$, the comparison group includes those unemployed who either never participate in any program or who start some treatment after month

3 (see table 2). Similarly, for treatment during strata 2 and 3, the basic sample consists of those individuals in group k and l who are still unemployed at the beginning of the stratum.

We then combine kernel matching involving the propensity score with cell matching by imposing additional criteria along which treated and comparison observations are aligned exactly. Specifically, we impose the following four matching requirements: (i) similarity in the pairwise propensity score, (ii) similarity in the calendar date of the beginning of unemployment, (iii) equality in the elapsed time in open unemployment, and (iv) equality in the employment history before the start of unemployment. Given exact matches for iii and iv, we estimate the counterfactual employment and earnings outcomes by means of a Nadaraya-Watson kernel regression to achieve i and ii. We use a product kernel in the estimated propensity score and the calendar month of entry into unemployment,

$$KK_j(p_d, c_d) = K\left(\frac{p_d - p_j}{h_p}\right) \times h_c^{|c_d - c_j|}, \tag{3}$$

where $K(z)$ is the Gaussian kernel function; p_d and c_d are the propensity score and the calendar month of entry into unemployment of a particular treated individual d; and p_j and c_j are the estimated propensity score and the calendar month of entry into unemployment of an individual j belonging to the comparison group of individuals treated with l. The parameters h_p and h_c are the bandwidth (type) parameters. We use the bivariate cross validation procedure suggested in Fitzenberger et al. (2008) and Bergemann, Fitzenberger, and Speckesser (2009) to obtain the bandwidths h_p and h_c by minimizing the squared prediction error of the average l-outcome for the individuals in the l-group who are most similar to the participants in program k.[19] We obtain standard errors and pointwise confidence bands for our estimated treatment parameters through bootstrapping based on 250 resamples. We resample individuals with their entire observation vector to obtain standard errors clustered at the individual level, thus also accounting for the multiple appearance of individuals. The propensity score is reestimated in each resample.

Recently, there have been a number of Monte Carlo studies analyzing the performance of various estimators for treatment effects relying on propensity score matching (see Galdo, Smith, and Black 2008; Busso, DiNardo, and McCrary 2009, 2011; and Huber et al. 2010). These studies reassess the conclusions drawn by the earlier Monte Carlo study of Frölich (2004). The key insights from the different studies are the following. First, there is no estimator that performs best in all settings and according to all criteria. Second, the popular nearest-neighbor matching approach is inferior to more in-

[19] For the comparisons of training versus searching, the cross validation yields an optimum for h_c of zero, which corresponds to an exact alignment in calendar time.

volved estimators appropriately adjusted to the estimation problem investigated. Third, optimized versions of both kernel or radius matching improve upon simple implementations, unless sample sizes are really small, and they typically do not perform badly even when inverse probability weighting (IPW, the preferred method in Busso et al. [2009, 2011]) performs better. Fourth, regression adjustment in addition to the preferred estimator by the authors improves the performance when the propensity score is incorrectly specified.[20] The aforementioned Monte Carlo studies determine the bandwidth for kernel matching approaches based on a standard leave-one-out cross validation. The Monte Carlo results reported in Galdo et al. (2008) show that it can be useful to base the bandwidth choice on predictions for nontreated individuals who are close to the treated individuals.

Our estimation approach is not covered explicitly by the recent Monte Carlo studies reported in Busso et al. (2009, 2011) and Huber et al. (2010). Both local linear ridge matching (the best performing estimator in Frölich [2004] and Huber et al. [2010]) and IPW are not feasible in our application because we combine cell matching and kernel matching and match on closeness in two dimensions at the same time. In the case of treatment versus searching, we have sizable treatment groups (between 200 and 900 observations) and 17–75 times larger nontreatment groups (see table 2), and we solely estimate the average treatment effect on the treated. Furthermore, the overlap in the covariate distributions and the balancing of the covariates are both very good in these cases (see Sec. V.C). We expect our kernel regressions to perform very well in the large nontreatment samples, in particular because we take great care in details such as bandwidth choice and regression adjustments after matching (see Sec. IV.F). We follow Huber et al. (2010) in that we match exactly on employment history variables and in that we investigate the importance of trimming highly influential comparison observations (see Sec. V.C).

C. Empirical Content of the DCIA

The DCIA in equation (2) states that conditional on a given unemployment experience and a vector of observed covariates, the sequence of potential outcomes associated with alternative treatment l (including nonparticipation $l = 0$) in the same stratum is mean independent of the treatment status in this stratum. In a dynamic context, nonparticipation in the current stratum entails the possibility of participation in later strata. Our matching approach will produce valid estimates if we consider all the determinants that jointly influence treatment status and potential outcomes in the current stratum as well as in future strata. Conditional on these determinants,

[20] Note that the studies do not analyze the effect of regression adjustment for all estimation approaches.

individuals are randomly allocated to one of the treatments in a given stratum, and there is balanced anticipation of future treatment or employment chances.[21] We argue in the following that these assumptions are plausible in light of program assignment in Germany and the rich information in our data.[22]

Although the assignment to a training program typically occurs with the consent of the job-seeker, considering his or her willingness to receive training and to work in a specific field, the assignment decision is up to the discretion of the caseworker during the time period we consider. According to Blaschke and Plath (2000), indicators available to the caseworker, like the composition of a group of participants in a particular course or his or her assessment of the motivation of the unemployed, played an important role. Job placement has priority over program participation. The unemployed are encouraged to continue job search at any time, even while participating in a training program. The assignment to a particular training program is driven strongly by the supply of courses (Schneider et al. 2006). Evidence from qualitative surveys in Schneider et al. (2006) suggests that belated assignments and referrals on very short notice are commonly used in order to assure a high capacity utilization of courses booked in advance and to keep up job search incentives. This suggests that program assignment is not targeted but is driven by region-specific variation in the supply of courses throughout the year. From the perspective of the unemployed, the assignment to a particular program cannot be perfectly anticipated. Moreover, the Database of Registered Job-Seekers involves a wide range of information that is collected by the caseworker as a basis for his or her counseling activities and assignment decisions.

To be specific, we consider the following variables that reflect the caseworker's information on the motivation, plans, and labor market prospects of a particular unemployed: the caseworker's assessment of the job-seeker's current health status, information on his or her previous health status (during the last 2 years before the start of the current unemployment spell), a dummy variable indicating whether the unemployed person appeared to lack motivation (e.g., failed to show up at regular meetings), dummies indicating whether the job-seeker dropped out of a program, whether benefits were withdrawn, and whether the person participated in a program providing psychosocial support, where all variables refer to the last 3 years unless indicated otherwise. In addition, we include variables indicating whether the job-seeker would like to work in a different occupation,

[21] "Balanced" means that treated and nontreated have the same predictions conditional on the observables considered.

[22] See appendix section A for a list of the covariates considered and descriptive statistics. See Osikominu (2013) and her supplementary appendix for further institutional information.

whether he or she is looking for a part-time job, and the number of job proposals he or she received from the employment office.[23]

The literature (e.g., Heckman, Ichimura, and Todd 1997; Heckman et al. 1998; Heckman et al. 1999; Heckman and Smith 1999) stresses the importance of conditioning flexibly on lagged employment and wages, benefit receipt history, and local labor market conditions. Our sample of analysis involves individuals who, between 2000 and 2002, exit regular employment lasting 3 months or longer. Our data allow us to include in the propensity score a rich set of variables depicting past employment and nonemployment histories, as well as local labor market conditions. To account for potential anticipatory effects regarding reemployment, we use the complete longitudinal information in our administrative data sources covering employment and benefit receipt and impose several restrictions to ensure that our sample as a whole as well as the different treatment and comparison groups are sufficiently homogeneous. We use several different variables recording wages prior to unemployment as well as occupation and industry of the last job, indicators of whether the last job was part-time, whether it was a white-collar or blue-collar position, the reason why the job ended, and the quarter of the beginning of the unemployment spell. We also include indicators of whether the person was employed in month 6, 12, or 24 before the beginning of the unemployment spell, the number of days employed during the preceding 7 years, and the average annual earnings and the annual employment status in each of the 7 years before the beginning of the unemployment spell.

As in Card and Sullivan (1988), we stratify treatment and comparison individuals based on their employment history sequences as well as elapsed unemployment duration (see also Dolton and Smith 2011). We construct 10 different employment history groups that indicate different combinations of the annual employment status of an individual during the 4 years preceding the unemployment spell.[24] We use nine sequences, (1000), (1001), (1010), (1011), (1100), (1101), (1110), (1111), and (0000). The sequence (1010), for example, represents employment during the first and the third years before entry into unemployment and nonemployment during the sec-

[23] Note that these variables reflect information that is collected in the course of the counseling process using standardized, closed-ended answer formats. For the three variables occchange, endlastjob, and family (see appendix section A), the share of missings is relatively high. Nevertheless, these variables often prove significant in the propensity scores. This likely reflects the fact that the missings are informative as well. For instance, job-seekers who are undetermined about their future occupation might choose not to answer the question.

[24] A person is considered as employed in a given year if he or she is employed for at least 50% of the days. The results are not sensitive to the exact choice of the cutoff point.

ond and the fourth years. The nine sequences considered so far cover over 95% of the cases. The remaining sequences are subsumed in a residual category. We include dummy variables for each sequence in the propensity score.

Turning to nonemployment, we include the entitlement period for unemployment benefits at the beginning of unemployment, as well as variables measuring the duration of unemployment transfer payments (i.e., unemployment benefits, unemployment assistance, or subsistence allowance) in days during the last 3 years, the number of days without any information in the data set (e.g., periods of self-employment or out of the labor force), indicators for whether somebody was disabled or unable to work at some point in time during the preceding 3 years, and a variable recording previous contacts with the employment office during the last 3 years. Finally, to model local labor market conditions, we use several different variables on the unemployment rate in the county of residence, indicators for the federal state, and indicators for the economic situation of the local labor market.

D. Specification of the Propensity Scores

Using all of the variables described in the previous section as possible regressors, we fit the propensity scores separately for each of the six evaluation subsamples (men/women, stratum1/stratum2/stratum 3) and each treatment comparison pair. In each case, we run an extensive specification search. The final specification was chosen based on which variables (according to institutional and economic knowledge) may drive the selection into programs, based on statistical significance of the variables included and based on the balancing tests described below. We usually start with a fairly general specification and drop variables that are grossly insignificant. This leads to satisfactory specifications in most cases. In the few cases in which the balancing condition fails, we modify the specification and reinclude covariates or add interactions in order to eliminate the covariate imbalance. We take particular care in ensuring that pretreatment employment and earnings outcomes are well balanced between treatment and comparison group members. The final specifications typically include between 20 and 35 covariates (see appendix section B.1 for the estimation results; the appendix is available online as a supplementary PDF).

E. Testing for Covariate Balance

We employ two different balancing tests to check whether our treatment and comparison groups are sufficiently balanced.[25] First, we test whether the means of important covariates differ significantly between treatment and

[25] See Lee (2013) for a recent critical assessment of balancing tests.

matched comparison groups using bootstrap *t*-tests and Wald tests (for groups of variables that belong together). Our bootstrap approach accounts for clustering at the individual level, which is of particular importance as members of the comparison group can be used as comparisons more than once. In each subsample, the comparison of means is based on the same set of 90 regressors that we consider particularly important (see the discussion in Sec. IV.C). We calculate the matched mean of a given regressor by applying the matching procedure described above to this regressor in exactly the same way as we do to predict the counterfactual employment and earnings outcomes. Second, we investigate whether treated and matched nontreated individuals differ significantly in their employment and earnings outcomes over a period of 7 years before the beginning of unemployment. We estimate these differences in the same way as the treatment effects after the beginning of the program. By construction, treated individuals and their matched counterparts exhibit the same unemployment duration until the beginning of treatment as well as the same type of 4-year employment history before the beginning of the unemployment spell. We report the results of both balancing tests in Section V.C and in appendix section B.2 in the online appendix.

F. Regression Adjustment and Test for Effect Heterogeneity

The estimated ATT (average treatment effect on the treated) may be biased if treated and matched comparison individuals are not matched well by observed characteristics affecting the treatment effect.[26] Following Mueser et al. (2007) and Abadie and Imbens (2011), we apply a regression-based adjustment to investigate the existence of any remaining mismatch between treated and comparison individuals after kernel matching. While Abadie and Imbens (2011) propose the ex post regression adjustment as a device to remove the leading bias term of the nearest-neighbor matching estimator (Abadie and Imbens 2006), in our case of kernel matching, the bias of the nonparametric kernel regression vanishes asymptotically. Nevertheless, as some mismatch in the levels of the matching variables between each treated unit and the comparison units may remain in any finite sample (Mueser et al. 2007, 765) we use the ex post regression adjustment to investigate the existence of such mismatch bias.[27] At the same time, we use it to test for treat-

[26] A possible mismatch may be caused by a misspecification of the propensity score or by the fact that, despite the validity of the DCIA, the particular sample draws for the observed characteristics may differ between treated and matched comparison individuals. The former involves a bias that remains in a large sample if a parametric propensity score is used, while the latter involves a finite sample bias.

[27] Mueser et al. (2007) adjust each set of comparison cases for a particular treated case using a linear regression in the weighted data set of comparison cases (see also Heinrich et al. [2010] for this approach). Unlike us, they do not use the ex post regression to study effect heterogeneity.

ment effect heterogeneity.[28] In fact, estimation of the ATT provides a semiparametric, aggregate impact measure of potentially heterogeneous individual treatment effects depending upon observed characteristics of the participating individuals.

In our outcome regressions, the dependent variable, $\bar{Y}_i(u)$, is the average of the individual-period-specific outcomes over 24 months from month u onward. We focus on this global outcome in order to abstract from the volatility over time of the monthly outcomes. Pooling over all possible starting dates u in stratum s, that is, $u \in [\underline{u}_s, \bar{u}_s]$, $s = 1, 2, 3$, we run the following weighted regression:

$$\min_{\hat{\alpha},\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\delta}}} \sum_i \sum_{u=\underline{u}_s}^{\bar{u}_s} g_i(u)[\bar{Y}_i(u) - \hat{\alpha} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \hat{\gamma}D_{s,i} - D_{s,i}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})\hat{\boldsymbol{\delta}}]^2, \qquad (4)$$

where the summation is over all individuals i who either receive treatment k or l in the stratum considered, $\boldsymbol{X}_i$ denotes a vector of observed characteristics, $\bar{\boldsymbol{X}}$ denotes the average of the characteristics in the sample of individuals receiving treatment k,[29] and $D_{s,i}$ is a dummy variable for receiving treatment k in stratum s, that is, $D_{s,i} = 1$ ($D_{s,i} = 0$) if individual i receives treatment k (l).

The variable $g_i(u)$ is the weight associated with treatment start u. For a treated individual ($D_{s,i} = 1$), the weight picks the observed start month, that is, $g_i(u) = 1$ for persons starting treatment k in month u and zero otherwise. For a nontreated individual j ($D_{s,j} = 0$) serving as a comparison for treatment k starting in month u, the weight $g_j(u)$ equals

$$\sum_{d \in \{T_u = k \wedge U^o = u-1; u \in [\underline{u}_s, \bar{u}_s]\}} \frac{K[(p_d - p_j)/h_p] \times h_c^{|c_d - c_j|}}{\sum_{n \in \{T_{\tilde{u}} = l \wedge U^o \geq u-1; u, \tilde{u} \in [\underline{u}_s, \bar{u}_s]\}} K[(p_d - p_n)/h_p] \times h_c^{|c_d - c_n|}},$$

where d indexes those individuals starting treatment k in month u ($T_u = k \wedge U^o = u - 1$) and n those comparison units who are still in open unemployment in the month before treatment k starts ($U^o \geq u - 1$). The function inside the sum corresponds to the normalized kernel weight as given in equation (3), where the normalization guarantees that the total sum of weights for d across nontreated individuals j sums to one. Note that j is used repeatedly for each possible treatment starting date u in the

[28] To our knowledge, recent evaluation studies rarely estimate ex post outcome regressions to study effect heterogeneity. In the literature on statistical treatment choice, Frölich (2008) uses a similar approach to predict potential outcomes as a function of individual characteristics. Huber et al. (2010) report Monte Carlo evidence for treatment effect estimators based on such ex post outcome regressions, but they do not investigate effect heterogeneity.

[29] All the explanatory variables in $\boldsymbol{X}_i$ are dummies.

respective strata for which j serves as a comparison unit, that is, observation j is used up to $(\bar{u}_s - \underline{u}_s + 1)$ times in the pooled regression (4).

The term $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ in equation (4) controls for the effect of the characteristics on the average outcome variable. The coefficient on the treatment dummy, $\hat{\gamma}$, gives the mismatch-corrected estimate of the aggregate ATT. This mismatch-corrected estimate can be compared to the estimate obtained without the additional regression adjustment. If the two estimates are similar, then our matching procedure works well in aligning the distributions of the conditioning variables in the comparison group to that of the treatment group. The coefficients on the vector of interaction terms, $\hat{\boldsymbol{\delta}}$, give the ceteris paribus changes in the treatment effect when changing the corresponding characteristics. If $\boldsymbol{\delta} = 0$, then there is no effect heterogeneity by the level of covariates. The standard errors of the estimated coefficients in (4) are obtained through a full bootstrap for the matching estimator by rerunning the regression for all resamples.

V. Which Program Works and for Whom?

A. Effectiveness of Training Compared with Searching

Figures 3–10 show the evaluation results for the comparison of training versus not participating in any active labor market program (i.e., "searching") in a given time window. We distinguish three different time windows (strata) of elapsed unemployment experience: 1–3 months (stratum 1), 4–6 months (stratum 2), and 7–12 months (stratum 3). Each set of graphs (e.g., figs. 3 and 4 for males and females; hereafter denoted as figs. 3/4) displays, for different points in time given on the horizontal axis, the average treatment effect on the treated (ATT), that is, the difference between the observed outcome with training and the estimated counterfactual outcome without training averaged over those who participate in the program in a given time window. On the time axis in our graphs, positive values denote months since program start, while negative values represent preunemployment months and years. We omit the period between the start of unemployment and the start of the program in which both comparison and treatment persons are unemployed. The dashed lines around the estimated ATT are bootstrapped 95% pointwise confidence bands. Treatment effects for a particular time period are statistically significant if zero is not contained in the confidence band. While figures 3/4 and 7/8 refer to the difference in employment rates, figures 5/6 and figures 9/10 represent differences in pretax earnings. For both outcome measures, subsidized employment is counted as nonemployment.[30]

[30] Our earnings information comes from social security records, which is collected only once in a given calendar year if a person is employed with the same employer throughout and no changes in relevant information, such as a change of

Turning to the estimated employment effects of short-term training (STT) in figures 3/4, we find negative effects shortly after the program start in the order of −5 percentage points (ppoints). This suggests that, while enrolled in the program, participants have a 5 percentage point lower monthly employment rate than they would have if they did not participate in the program. These lock-in effects do not last for more than 2–3 months, which is in line with the short duration of STT programs of about 1 month. After the lock-in period, the treatment effects turn positive in general, but they are not always statistically significant. The results differ strongly across strata. While there is no evidence for statistically significant treatment effects for individuals participating in the first 3 months of their unemployment spell (stratum 1), treatment effects for men starting a STT program in months 7–12 of their unemployment spell (stratum 3) and women starting one in month 4 or later are significantly positive. Accordingly, the monthly employment rate of men (women) participating in STT is increased by about 5 percentage points (about 11 percentage points) after the end of the lock-in period (see also table 3). The impact estimates for earnings in figures 5/6 show a corresponding pattern. There are statistically significant earnings gains of up to €100–€200 for female participants in STT in strata 2 and 3 and marginally significant effects of up to €100 for male participants in stratum 3 and female participants in stratum 1. Averaging over the second year after program start, the monthly effect is €87 for men in stratum 3 and €169 for women in strata 2 and 3 (see table 4).

Estimates of the employment effects for the more substantive classroom further training programs (CFT) are given in figures 7/8. The most conspicuous difference between these results and those for STT programs is the long and pronounced lock-in effect. During the first 6 months after program start, participants have an employment rate that is up to 25 percentage points lower than it would have been if they had not taken part in the program. From month 7 onward, the treatment effects recover, eventually turning positive 10–12 months after program start. The lock-in period is longer and entails larger employment losses for people who take up their treatment during the first 6 months of unemployment (strata 1 and 2). In contrast, people who have been unemployed for more than 6 months (stratum 3) experience less deep and shorter lock-in effects. While there is little evidence for statistically significant employment effects for men starting a CFT program in months 1–6 of their unemployment spell (strata 1 and 2) or women enrolling in the first 3 months of unemployment (stratum 1), treatment effects for longer-term unemployed men (stratum 3) and medium-to longer-term unemployed women (strata 2 and 3) are sizable and

the health insurance, occur within the year. Observed changes in monthly earnings are therefore due to wage changes with the same employer, job-to-job changes, or job-to-nonemployment transitions.

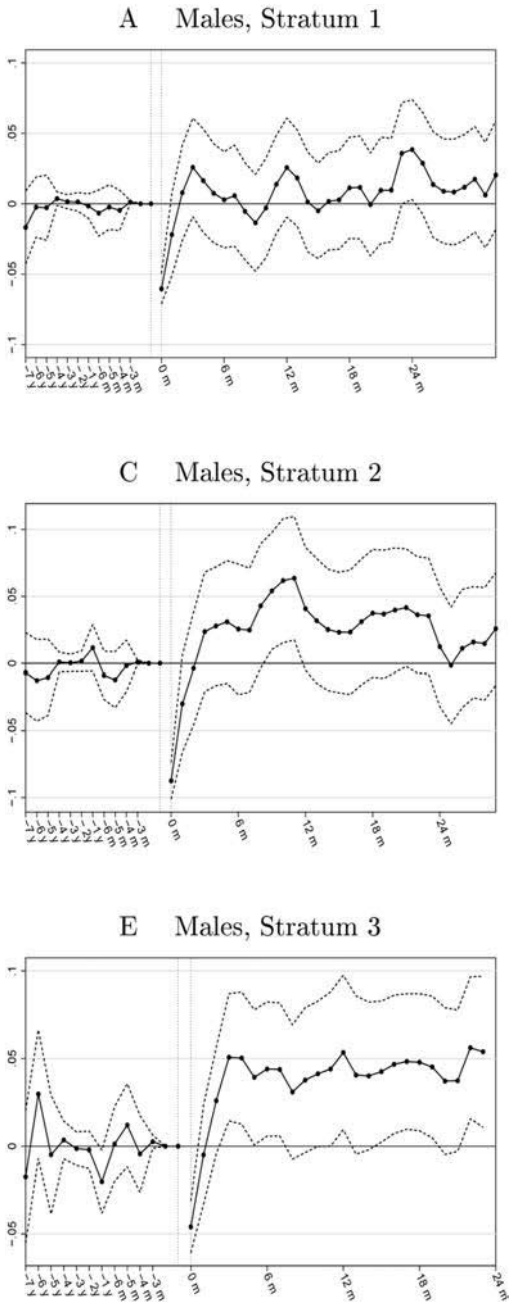

FIG. 3.—Employment effects for STT versus searching, males. Difference in employment rates is measured on the ordinate, preunemployment (< 0) and posttreatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence intervals. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and stratum 3 during months 7–12.

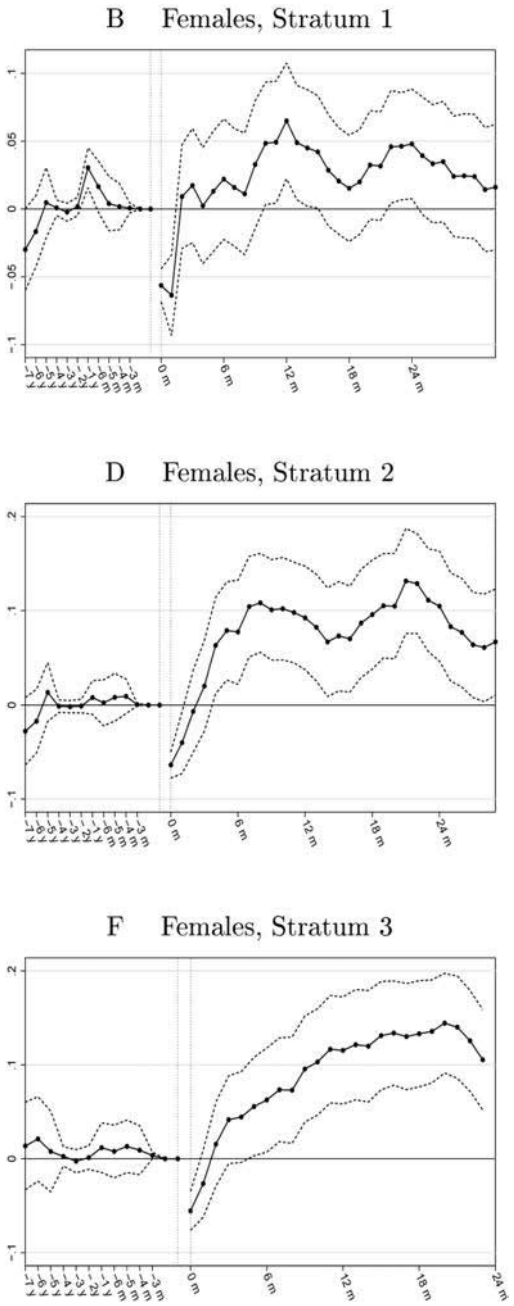

FIG. 4.—Employment effects for STT versus searching, females. Difference in employment rates is measured on the ordinate, preunemployment (< 0) and posttreatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence intervals. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and stratum 3 during months 7–12.

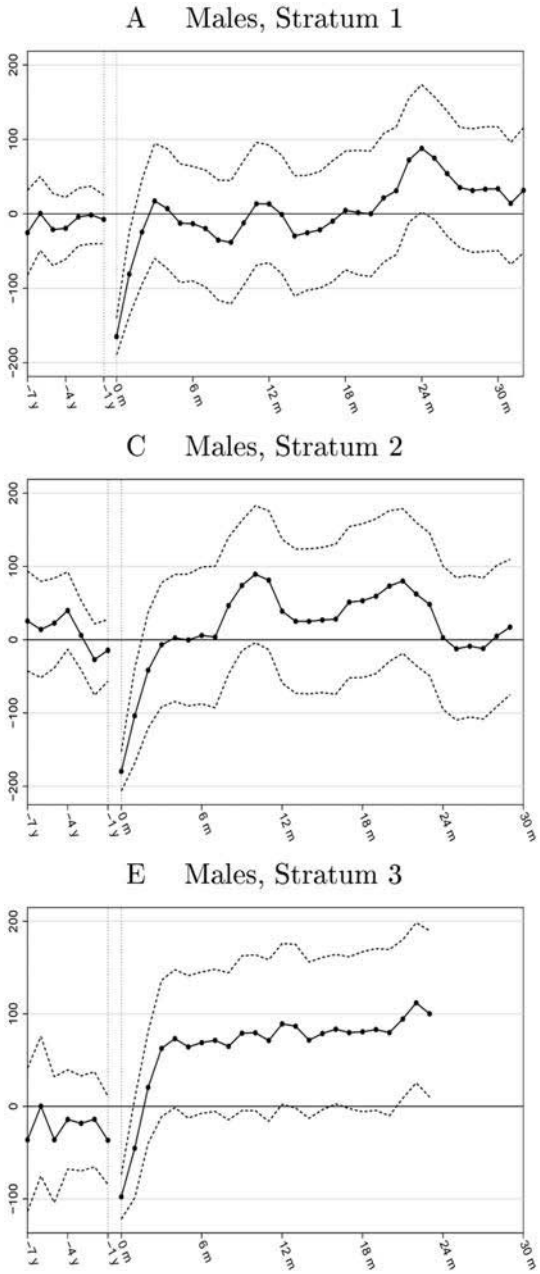

FIG. 5.—Earnings effects for STT versus searching, males. The abscissae measures preunemployment (< 0) and posttreatment (≥ 0) months and years. The difference in average real earnings (in euros, reference year 2000) during the corresponding time interval is given on the ordinate. Dotted lines are pointwise 95% bootstrap confidence intervals. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and stratum 3 during months 7–12.

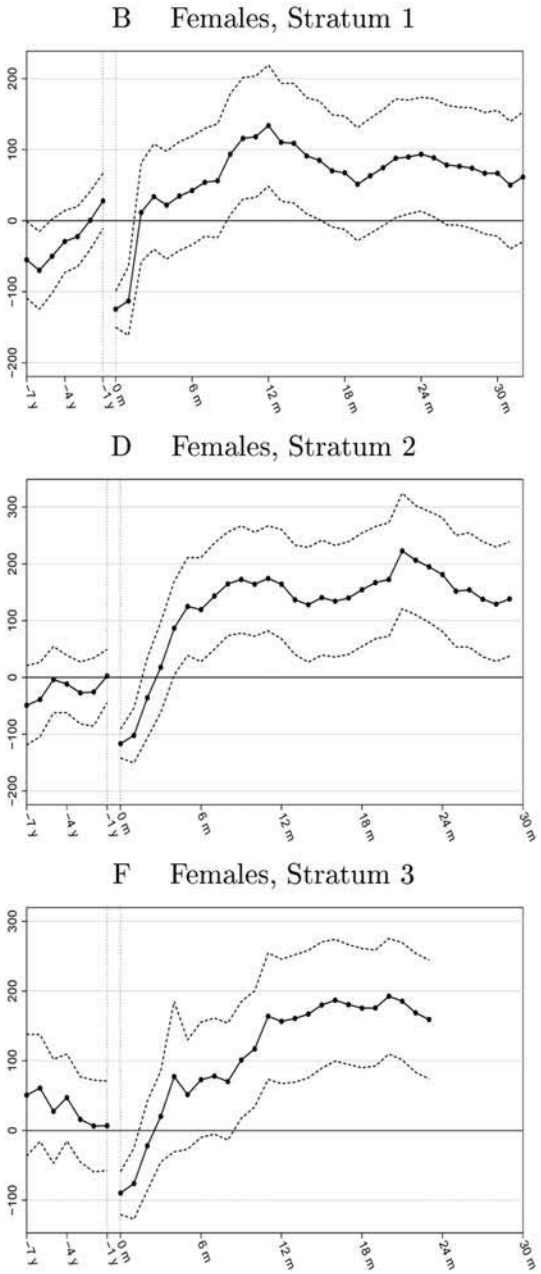

FIG. 6.—Earnings effects for STT versus searching, females. The abscissae measures preunemployment (< 0) and posttreatment (≥ 0) months and years. The difference in average real earnings (in euros, reference year 2000) during the corresponding time interval is given on the ordinate. Dotted lines are pointwise 95% bootstrap confidence intervals. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and stratum 3 during months 7–12.

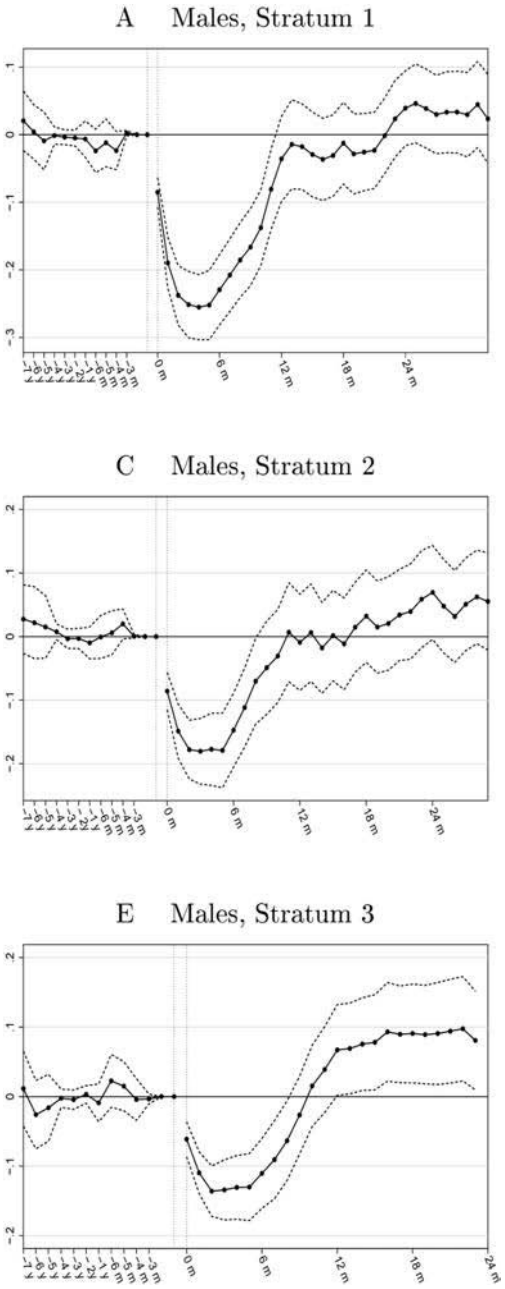

FIG. 7.—Employment effects for CFT versus searching, males. Difference in employment rates is measured on the ordinate, preemployment (<0) and posttreatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence levels. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and, stratum 3 during months 7–12.

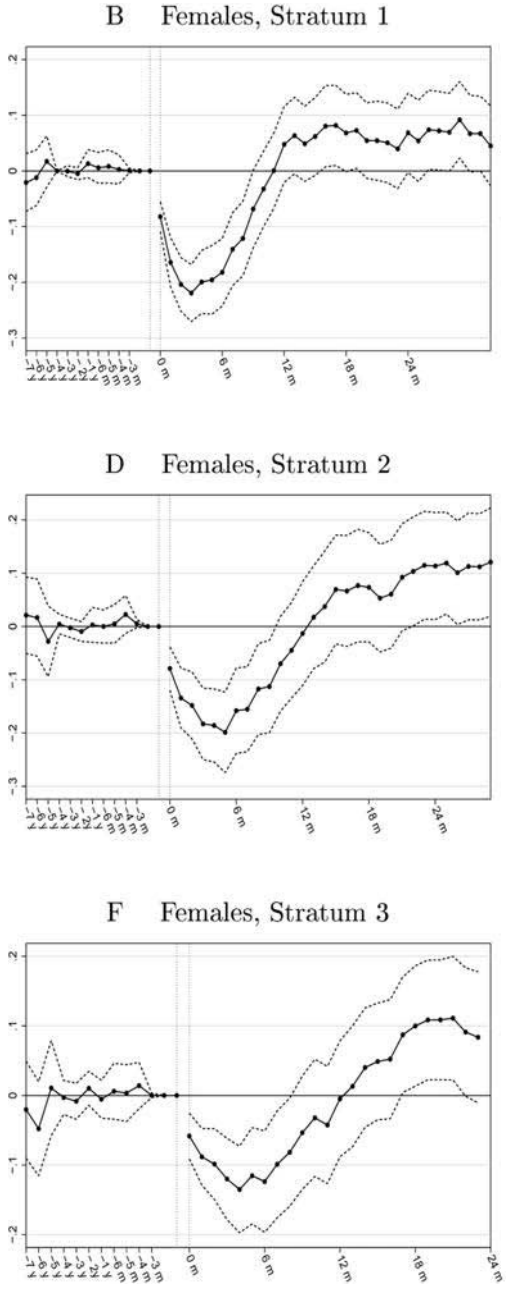

FIG. 8.—Employment effects for CFT versus searching, females. Difference in employment rates is measured on the ordinate, preemployment (< 0) and posttreatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence levels. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and stratum 3 during months 7–12.

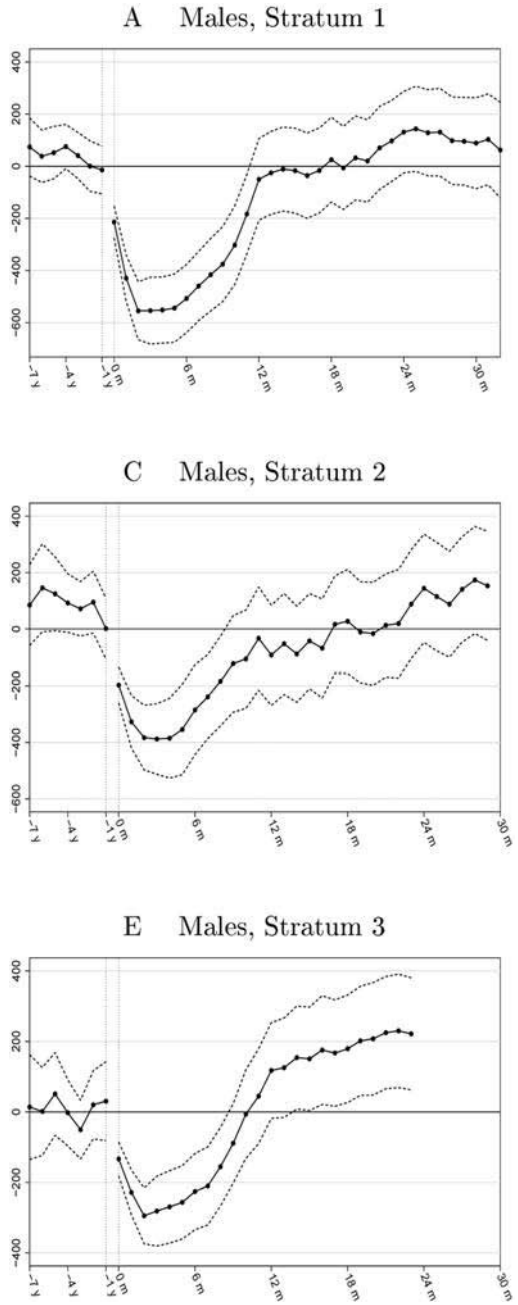

FIG. 9.—Earnings effects for CFT versus searching, males. Difference in employment rates is measured on the ordinate, preemployment (< 0) and posttreatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence levels. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and, stratum 3 during months 7–12.

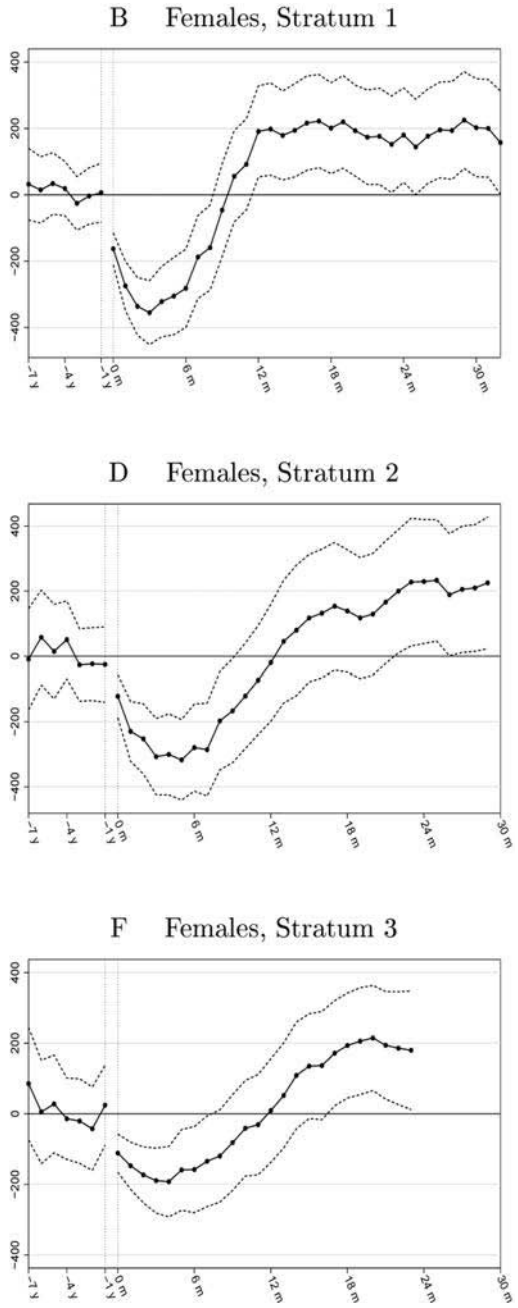

FIG. 10.—Earnings effects for CFT versus searching, females. Difference in employment rates is measured on the ordinate, preemployment (< 0) and post-treatment (≥ 0) months and years on the abscissa. Dotted lines are pointwise 95% bootstrap confidence levels. Stratum 1 denotes entry into program during months 1–3 of unemployment, stratum 2 during months 4–6, and, stratum 3 during months 7–12.

Table 3
Average Monthly Treatment Effects on the Treated: Employment

	Stratum 1		Stratum 2		Stratum 3	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Men:						
STT versus searching	−.002	.010	.019	.034*	.030**	.046**
	(.013)	(.016)	(.017)	(.020)	(.015)	(.018)
STT versus CFT	.163***	.021	.173***	.062	.149***	.057
	(.024)	(.038)	(.024)	(.041)	(.018)	(.035)
CFT versus searching	−.190***	−.019	−.113***	.015	−.078***	.084***
	(.018)	(.026)	(.023)	(.032)	(.017)	(.032)
CFT versus STT	−.120***	.028	−.100***	−.026	−.061**	.076*
	(.027)	(.037)	(.034)	(.047)	(.028)	(.044)
Women:						
STT versus searching	.008	.037**	.053***	.096***	.050	.128***
	(.016)	(.018)	(.019)	(.025)	(.021)	(.025)
STT versus CFT	.143***	0	.194***	.021	.137***	.067
	(.022)	(.041)	(.027)	(.056)	(.023)	(.047)
CFT versus searching	−.134***	.060*	−.132***	.063	−.087***	.070*
	(.022)	(.031)	(.029)	(.046)	(.027)	(.038)
CFT versus STT	−.141***	−.011	−.161***	−.012	−.100***	−.071
	(.028)	(.040)	(.036)	(.053)	(.038)	(.054)

NOTE.—STT = short-term training; CFT = classroom further training. Strata 1, 2, and 3 refer to programs starting in months 1–3, 4–6, and 7–12 of unemployment, respectively. Columns labeled "Year 1" refer to the average of the monthly treatment effects across the first 12 months since program start, and columns labeled "Year 2" refer to the average of the monthly treatment effects during the second year since program start. Bootstrapped standard errors based on 250 replications are in parentheses.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

statistically significant. After the initial lock-in phase, they eventually exceed 10 percentage points, with an average of 7 percentage points throughout the second year since program start (see also table 3).

Figures 9/10 reveal a similar pattern for the earnings impacts of CFT compared with searching. After an initial lock-in effect, there are significant earnings gains for male participants in CFT in stratum 3 (and marginally in stratum 2) and in all three strata for women. Especially for men, the earnings gains after lock-in tend to be larger than those for STT, often attaining €200 per month. This means that although after lock-in CFT and STT yield comparable employment gains, CFT participants benefit on average more from their program than participants in STT, which is consistent with the fact that CFT involves a more sizable human capital investment compared to STT. Hence, CFT programs, unlike STT programs, contribute to the development of human capital that is rewarded on the labor market.

Tables 3 and 4 summarize the estimated treatment effects during the first 2 years after program start. We report the difference of the estimates regard-

Table 4
Average Monthly Treatment Effects on the Treated: Earnings

	Stratum 1		Stratum 2		Stratum 3	
	Year 17	Year 2	Year 1	Year 2	Year 1	Year 2
Men:						
STT versus searching	−30.382	4.693	−2.509	47.598	42.626	86.533**
	(30.723)	(37.668)	(34.285)	(46.249)	(30.476)	(40.003)
STT versus CFT	314.914***	−55.558	300.415***	160.484*	279.830***	123.143*
	(54.905)	(97.928)	(63.698)	(92.818)	(37.369)	(70.515)
CFT versus searching	−424.437***	6.675	−250.291***	−17.567	−175.549***	179.112**
	(51.135)	(72.886)	(58.449)	(81.542)	(38.039)	(70.706)
CFT versus STT	−269.484***	139.711	−193.397**	−142.781	−78.360	220.078**
	(59.720)	(91.966)	(75.756)	(111.560)	(56.994)	(99.181)
Women:						
STT versus searching	28.555	86.124**	76.040**	163.312***	46.941	174.039***
	(30.098)	(38.009)	(33.035)	(45.979)	(31.321)	(40.091)
STT versus CFT	257.059***	−30.683	307.633***	34.625	177.653***	54.631
	(40.874)	(83.324)	(46.511)	(102.059)	(36.686)	(79.961)
CFT versus searching		193.081***				
	−190.229***	(65.128)	−220.986***	123.872	−128.160***	148.351**
	(45.236)		(53.101)	(88.116)	(45.702)	(69.050)
CFT versus STT	−250.685***	−2.396	−250.348***	4.981	−107.283*	−11.134
	(59.480)	(85.558)	(63.550)	(101.255)	(57.787)	(97.949)

NOTE.—STT = short-term training; CFT = classroom further training. Strata 1, 2, and 3 refer to programs starting in months 1–3, 4–6, and 7–12 of unemployment, respectively. Columns labeled "Year 1" refer to the average of the monthly treatment effects across the first 12 months since program start, and columns labeled "Year 2" refer to the average of the monthly treatment effects during the second year since program start. Bootstrapped standard errors based on 250 replications are in parentheses.

* $p < .10$.
** $p < .05$.
*** $p < .01$.

ing the average monthly employment and earnings gains or losses over the course of the first year and the second year. For STT versus searching (rows 1 and 5 of tables 2, 3, and 4), in cases with slightly negative effects during the first year, these are generally offset by positive effects in the second year. Thus, the average effect tends to be zero or positive for STT after 2 years. In contrast, it remains negative in the majority of cases for CFT versus searching after 2 years (rows 3 and 7 of tables 2, 3, and 4). Consistent with the graphical results, when we compare the estimates across strata, the estimated effect increases (becomes less negative) the later a STT or CFT program starts during unemployment.

What explains the differing treatment effects across strata? First, a large number of those having become unemployed recently find new jobs easily without participating in a training program. If these people are assigned to a CFT program anyway, they will be "locked-in," while many of their counterparts in the comparison group quickly exit to employment. The evidence in Osikominu (2013) suggests that the opportunity cost of participating in a long program is in fact higher for people with a high probability of exiting unemployment without treatment. Our findings in Section VI are also consistent with this view, because early program participants tend to

be a positive selection of all unemployed and are therefore matched with nonparticipants who exit unemployment very quickly. Second, treatment effects tend to be higher for the late program starts if many of the long-term unemployed in the comparison group abandon their job search and move out of the labor force. Our additional evaluation results for the outcome variable "leaving the labor force" (see appendix section C.1 in the supplemental online appendix) confirm that participants enrolling after a longer unemployment experience have a reduced probability to drop out of the labor force compared to the situation if they did not receive training. Hence, an important channel through which training programs work consists in keeping the long-term unemployed in the labor force.

B. Heterogeneity of Training Impacts

Given that in many cases training programs show positive treatment effects in the medium run when compared to attending no program in the same stratum, the questions arise: (i) Who benefits most from a given program? (ii) Which program is the best for a given subpopulation? The results in the previous section suggest that STT may have similar or even more positive effects compared to CFT when each of the programs is compared to attending no program. However, this does not necessarily mean that participants in STT could not have improved their employment chances by attending CFT instead or that participants in CFT would not have lost from taking part in STT instead. This is because the selection of the treated may differ by treatment.

Tables 3 and 4 also contain estimates for the direct comparisons between STT and CFT, which allows us to address the second question (see also appendix section C.2 in the supplemental online appendix for further evidence). Row 2 for men and row 6 for women in tables 2 and 3 show the average effect of participating in STT instead of CFT for the STT participants. The estimates are generally positive, suggesting that those participating in STT could not have fared better with the more intensive CFT program. For instance, the numbers in row 2, columns 1 and 2 of table 2, suggest that male participants in STT in stratum 1 have on average a 16 percentage point higher employment rate in year 1 and a 2 percentage point higher employment rate in year 2 compared to the situation in which they would participate in CFT. The advantage of STT over CFT for those who participate in STT is generally higher in year 1 than in year 2, which reflects the differing enrollment lengths of the two programs. When we look at the opposite comparison of CFT versus STT for those who participate in CFT (rows 4 and 8 in tables 2 and 3), similar patterns emerge during the first year after program start. During the second year after program start, when most CFT participants have completed their training, the estimated effect changes in favor of CFT. Overall, however, CFT participants could in most cases have done similarly well or even better with a STT program. This is re-

markable since it means that these people could have been assigned to the much less expensive STT programs without deteriorating their labor market chances (although an overall assessment would also have to consider how long exactly the impacts of STT and CFT persist).

Now, we turn to the question of who benefits most from a given program. The results presented so far reveal some heterogeneity in the effects for different subpopulations. A consistent finding is that programs are only effective for those individuals who start a program after having been unemployed for some time. However, note that this is subject to the caveat that the selection of eligible individuals changes over the course of the unemployment spell. Another finding is that training effects are generally larger for women than for men (figs. 3/4 and figs. 7/8). In the following, we investigate whether the treatment effect depends upon further observable personal characteristics. For this purpose, we regress the individual-level outcomes averaged over the first 24 months since program start on the treatment dummy, a set of observed covariates, and the interactions with the treatment dummy. The comparison observations are weighted according to their matching weights (see Sec. IV.F). We estimate a first regression including all the personal characteristics and interactions with the treatment dummy and three further regressions, where we only include interaction effects for one out of three subsets of characteristics (i.e., for demographic variables, human capital variables, and characteristics of the previous job). By adding the interaction effects separately, the power of the tests is likely to be higher.

The results of the Wald tests for effect heterogeneity for all interaction variables are reported in table 5 and for the separate groups in appendix sections D.2 and D.3 in the supplementary online appendix. The full coefficient estimates are reported in appendix section D.1. In almost all cases, the Wald test statistics for the null hypothesis that the interactions between the treatment dummy and the characteristics have zero coefficients are nowhere close to being significant, neither when the specification includes interactions with all individual characteristics nor when we undertook separate analyses by type of variables. The lowest p-values are around 10% in two cases, and they are much higher in the remaining 34 cases. A small number of single coefficients are significant, but no clear patterns emerge. The overall insignificance suggests that we should not interpret the single significant coefficients. We conclude for the comparisons of training versus searching that there is little heterogeneity along observed characteristics after conditioning on gender and stratum.[31]

[31] We are not aware of state-of-the-art evaluation studies that investigate effect heterogeneity within subgroups defined by gender and elapsed unemployment. Thus, it is difficult to relate our findings to the literature. Most closely related to our

Table 5
Results of Outcome Regressions on Treatment Dummy, Personal Characteristics and Interactions

	Stratum 1	Stratum 2	Stratum 3
STT, males:			
χ^2 (p-value)	14.95 (.780)	22.02 (.339)	19.75 (.473)
Corrected ATT with CI	.005 [−.021, .030]	.025 [−.009, .060]	.037 [.007, .067]
Uncorrected ATT (SE)	.004 (.013)	.026 (.017)	.037 (.015)
STT, females:			
χ^2 (p-value)	23.81 (.251)	16.02 (.715)	20.38 (.434)
Corrected ATT with CI	.021 [−.010, .052]	.073 [.033, .112]	.089 [.047, .131]
Uncorrected ATT (SE)	.023 (.016)	.074*** (.020)	.088*** (.021)
CFT, males:			
χ^2 (p-value)	16.22 (.703)	26.43 (.152)	15.51 (.747)
Corrected ATT with CI	−.109 [.148, .070]	−.054 [−.100, −.007]	−.006 [.050, .038]
Uncorrected ATT (SE)	−.104*** (.019)	−.050** (.025)	.002 (.022)
CFT, females:			
χ^2 (p-value)	8.10 (.991)	21.96 (.343)	15.64 (.739)
Corrected ATT with CI	−.038 [−.086, .010]	−.036 [−.102, .031]	−.016 [−.075, .042]
Uncorrected ATT (SE)	−.036 (.024)	−.035 (.033)	−.010 (.029)

NOTE.—Strata 1, 2, and 3 refer to programs starting in months 1–3, 4–6, 7–12 of unemployment, respectively. The table summarizes the main results from a regression of the average treatment effect on the treated (ATT) averaged over the first 24 months after treatment start on the treatment dummy, a set of personal characteristics, and interactions of the treatment dummy and the personal characteristics; see eq. (4). Rows labeled χ^2 (p-value) show the test statistic (p-value) of a Wald test of joint significance involving the interactions with the treatment dummy ($H_0 : \delta = 0$ in eq. [4]). Rows labeled Corrected ATT with CI show the coefficient on the treatment dummy (γ in eq. [4]) that corresponds to the estimated mismatch-corrected ATT and its confidence interval (CI). For comparison, the ATT without regression correction and its standard error are shown in the rows labeled Uncorrected ATT. All inference is based on 250 bootstrap resamples. For additional results on effect heterogeneity, see appendix section D.

* $p < .10$.
** $p < .05$.
*** $p < .01$.

What explains the generally higher treatment effects for women compared with men? While our finding is in line with the literature (see Bergemann and van den Berg [2008] for an overview), this view has been challenged in a recent study by Lechner and Wiehler (2011), who, using Austrian data, argue that positive effects might be overstated if fertility responds to

analysis is the paper by Lechner et al. (2011). They stratify the treatment samples for men and women in a number of dimensions and find some differences beyond gender differences. According to our reading of the paper, Lechner et al. (2011) do not seem to provide evidence for significant effect heterogeneity within subgroups defined by gender and the elapsed duration of unemployment.

prolonged unemployment. We cannot rule out such an effect because we do not observe pregnancies in our data. However, as indirect evidence against the hypothesis of an effect due to pregnancies, we find that treatment effects do not differ significantly between women below age 40, who are in their fertility phase, and women above age 40, who are mostly past their fertility phase. Other studies (e.g., Lechner et al. 2007; Osikominu 2013) have attributed part of the gender differences in training impacts to differences in the occupational composition. We could not detect heterogeneous effects across occupations and industries after conditioning on gender and stratum. Our additional results for the outcome indicating a permanent exit from the labor force (see appendix section C.1) suggest that the labor market attachment effect of training is slightly stronger for women than for men. This difference may partly explain the gender gap in employment and earnings impacts. Unfortunately, we cannot explore this issue further because our data do not allow us to determine exactly whether (and for what reason) an exit from the labor market occurred.

C. Assessing the Quality of Our Benchmark Estimates

We provide extensive checks of the quality of our benchmark matching approach, for which the estimates are reported above. First, we investigate the balancing of covariates by comparing the means of 90 covariates between treated and matched comparison groups. The detailed results are contained in appendix section B.2 in the supplemental online appendix. The rows labeled Benchmark refer to our benchmark matching procedure, rows labeled Raw to the raw, unmatched samples. We conduct *t*-tests (or Wald tests) of equality of means in the treated and comparison samples for a given variable (or a group of variables). The figures in appendix section B.2 show that, in nearly all cases, the *p*-values of the Wald test statistics clearly exceed 5% after conducting our benchmark matching procedure. The Monte Carlo evidence reported by Lee (2013), which is, however, not based on clustered standard errors, suggests that balancing tests of the type used here tend to reject covariate balance too often. Thus, we are confident that covariates are well balanced in our benchmark estimates.

Second, we find that the distributions of the propensity scores and covariates on which we match exactly overlap sufficiently between treated and comparison groups (see appendix section B.3). For the comparisons of training versus searching, the smallest (largest) propensity score values in the treatment group are larger (smaller) than the minimum (maximum) in the comparison group in most cases (see appendix section B.3.1). In the few cases where this does not hold, the extrema in the treated sample do not exceed the corresponding extrema in the comparison sample by more than 5 percentage points. For the comparisons of one training program against another, the overlap is slightly worse but still sufficient. Thus, we do not need to trim any of the propensity score distributions of

the treated. Further, appendix section B.3.2 shows how many treated and comparison observations had to be dropped because of empty cells for any of the covariates on which we match exactly. For the comparisons of training versus searching, no treated observations are dropped. For the pairwise comparisons of training programs, between 1% and 3% of the treated observations are dropped, except for CFT versus STT, where 17% cannot be used.

Third, as a balancing test on the employment history, we compute placebo treatment effects also for the 7 years before the start of the unemployment spell considered. This is also similar to the preprogram test of Heckman and Hotz (1989), but note that we partly match on employment history. If our matching procedure works well, treatment effects should be zero in that period. In 23 out of the 24 cases depicted in figures 3–10, pretreatment outcomes of treatment and comparison groups do not differ significantly during any of the 7 years before the start of the current unemployment spell. A slight outlier is the earnings impacts of STT for women in stratum 1 (panel A of figs. 5/6), where treated individuals have significantly lower earnings in year 6 before the start of unemployment. Consequently, the earnings gain induced by the program may be even larger than the estimates suggest. Appendix section C.2 contains the graphical results for the comparisons of one training program against another. Here again, employment outcomes are not significantly different in any of the preceding 7 years in any of the graphs.

Fourth, as in Huber et al. (2010), we examine the existence and role of overly influential observations on our benchmark estimates in appendix section B.4. Only in one out of the 24 cases a comparison person has a weight that exceeds 4%, that is, for the case of STT versus CFT, males, stratum 3. In 10 cases—all of them concerning pairwise comparisons of training programs—the maximal weight exceeds 2%. The difference between the benchmark estimates and those obtained without the comparison observations with weights above 4% and 2%, respectively, is in general small. In terms of kernel matching weight, the influence of the most important comparison individuals is thus much smaller than reported in Huber et al. (2010). This is due to the fact that we align individuals by elapsed unemployment duration and not by actual and hypothetical program start dates, which implies a smaller pool of potential comparison units than in our case. Correspondingly, excluding (trimming) the most influential comparison individuals (even with a lower threshold than used in Huber et al. [2010]) has little effect on the estimates. Thus, in our benchmark estimates, we do not need to trim the most influential observations.

Finally, we investigate whether after matching there remains mismatch between treated and comparison units. We use the outcome regressions on the treatment dummy, personal characteristics, and their interactions with the treatment dummy to compute mismatch-corrected treatment ef-

fects (see Sec. IV.F). The coefficient on the treatment dummy gives the treatment effect net of any mismatch in personal characteristics that may have remained after matching. If our matching approach works well, the mismatch-corrected average treatment effects should coincide with uncorrected average treatment effects. The results reported in table 5 show that the uncorrected ATT lies in no case outside of the 95% confidence interval of the mismatch-corrected ATT. Thus, mismatch does not appear to be a problem.

VI. How Sensitive Are Impact Estimates to Data Features and Methodological Choices?

Given the richness of our data and the wide range of possible specification choices, it is a highly relevant question as to what extent data features and methodological choices influence the empirical results. The following analysis is therefore meant to contribute to the understanding of what matters for the design and outcome of an evaluation analysis. We start from our most comprehensive specification (whose results we reported in the previous section) as a benchmark. We then drop or modify certain specification features, for example, particular information in the data or the matching approach. We analyze the sensitivity of the results for employment in a given time period. The complete results of the sensitivity analyses are available in appendix section E in the supplemental online appendix. Here we summarize the range of variation that may result as a consequence of data or methodological changes.

A. Sensitivity Analysis 1: Employment History

In our first sensitivity analysis, following Card and Sullivan (1988), Heckman and Smith (1999), and Dolton and Smith (2011), we investigate how much difference it makes to condition on past labor market outcomes in a flexible way. Starting from our benchmark specification, we first drop the requirement of exact matching on 4-year history sequences. In a second step, we remove in addition the history sequence dummies from the propensity score. In the third step, the remaining 7-year history variables are dropped. The fourth step further removes all information on the benefit history. In a fifth step, we return to the benchmark specification but omit the requirement to exactly match on the elapsed unemployment duration within strata. Note that this fifth step is conceptually different from the first four steps as these omit information about the preunemployment period, while the fifth step omits information about the period after the beginning of unemployment.

The results of this exercise are shown in appendix section E.1. Omitting the requirement of exact matching on employment histories in general changes impact estimates only a little (appendix figs. E1/E2). In the case of "CFT vs searching for men in stratum 2" treatment effects are pulled

downward by a moderately small amount, suggesting that not controlling for employment sequences leaves too many relatively successful individuals in the comparison group. However, this result does not generalize to the other cases. More typical is the case of "SST vs searching for men in stratum 2," where not conditioning on employment sequences makes hardly any difference. If anything, the estimates are slightly pulled upward compared with the benchmark. The same is true if in addition the employment history dummies are dropped from the propensity score and the window of pre-unemployment information is reduced from 7 years to 3 years (appendix figs. E3/E4 and E5/E6). If, in addition, benefit-related information is dropped, the estimates also change very little (appendix figs. E7/E8).

By contrast, dropping the requirement of exact matching on elapsed unemployment duration in the benchmark specification considerably reduces estimated treatment effects in most cases, especially for individuals with long unemployment durations (appendix figs. E9/E10). The case "SST vs searching for men in stratum 3" involves the strongest change: positive treatment effects are completely eliminated. The explanation is that treated individuals tend to be incorrectly matched with comparison group members who have already found employment if one does not control for the elapsed unemployment duration at program start. This pulls down estimated treatment effects.

Apart from the result that matching on elapsed unemployment duration is important, our results suggest that conditioning flexibly on details of the employment history makes little difference (steps 1–4). This is in contrast to some of the results reported in Card and Sullivan (1988), Heckman and Smith (1999), and Dolton and Smith (2011). Unlike in these studies, our evaluation sample is already constructed in a way that strongly conditions on past labor market experience. If this basically eliminated the heterogeneity, it would not be surprising that the way in which we control for past employment information does not matter that much. Therefore, we further investigate the role played by the employment history variables in driving the selection into treatment in our application. First, employment- and benefit-related variables are in many cases significant determinants of the estimated propensity scores (see appendix section B.1).[32] Second, our detailed balancing tests reported in appendix section B.2 show that our treatment and the comparison samples significantly differ with respect to most employment and benefit variables before matching. Moreover, when

[32] An exception from this are the 4-year history dummies. Apart from the case of STT versus searching for women, the 4-year history dummies are insignificant in the estimated propensity score. Other employment variables such as the number of days employed or receiving benefits in any year before unemployment start are often significant.

we carry out those balancing tests for selected variants of our sensitivity analysis, we find that matching based on more parsimonious specifications (e.g., steps 2 and 4) fails to balance important employment and benefit variables. Therefore, it is even more remarkable that omitting the detailed information contained in steps 1–4 causes such little changes in the estimated treatment effects.

B. Sensitivity Analysis 2: Personal Characteristics

In our second sensitivity analysis, we vary the amount of personal information used for matching and the way in which it enters the propensity score. First, we investigate what happens if we omit the personality variables (i.e., information on whether the person took part in a program providing psychosocial support in the past, whether there were signs of a lack of motivation during previous unemployment spells, information on past penalties or past dropout from a program, the number of job proposals received, and the desire to change the occupation). In a second step, we omit in addition rich personal information that was included in the propensity score in the benchmark specification (i.e., information on household type, number and age of dependent children, marital status, information on health, disability, previous part-time employment, and reasons why the last job was ended). In a related but separate step 3, we compare the benchmark specification with a specification in which all variables are only used in a fixed, basic way without undertaking a separate specification search for each group. In step 4, we compare our benchmark specification to one in which we only include variables used in the evaluation study by Mueser et al. (2007), which represents a typical application in this field. In step 5, we drop the matching on the propensity score. In step 6, we omit in addition the exact matching on employment histories and align treated and comparison units only with respect to their prior unemployment experience and calendar time.

The results in appendix figures E11/E12 in the supplemental online appendix show that omitting the personality variables changes impact estimates surprisingly little as long as all other variables of the benchmark specification are controlled for. By contrast, omitting in addition other personal information may have a considerable effect (appendix figs. E13/E14), especially for early program starts (stratum 1). The direction of the changes is not entirely consistent, but omitting rich personal information tends to overstate treatment effects, suggesting that treated individuals generally have favorable personal characteristics. The fact that the inclusion of personality variables does not change much suggests that the kinds of information they represent are already largely included in the employment history variables and in the other personal characteristics. This is in contrast to the findings in Dolton and Smith (2011), where similar attitudinal variables

make a big difference.[33] The results for step 3 (appendix figs. E15/E16) suggest that varying the amount of specification search leads to relatively minor changes in most cases. The changes also do not display any clear direction.

Mueser et al. (2007) provide a recent state-of-the-art evaluation of job training programs based on administrative data from Missouri. Their setup is representative of what is possible using administrative data taken from different sources in the United States and other countries. Our data are also based on administrative registers, but they are substantially richer in terms of individual characteristics and details of employment histories. Step 4 compares our benchmark specification to a basic specification that is as close as possible to the one used in Mueser et al. (2007, 768): basic demographic and educational variables, that is, age, education, occupation, nationality, and region—as well as 4 quarters of earnings history and 4 quarters of employment dummies based on whether earnings in the given quarter were positive. We stick to our requirement to exactly match on the unemployment duration prior to treatment even though Mueser et al. do not measure unemployment durations at a monthly frequency. This implies that the treatment and comparison groups in Mueser et al. (2007) are more heterogeneous compared to our sample. Dropping this requirement has a strong impact on our results, as step 5 of sensitivity analysis 1 has shown, but it may be less important in the institutional environment in which program participation does not require unemployment.

The results for step 4 are shown in appendix figures E17/E18. In most cases treatment effects are overstated if matching is based on the more parsimonious information set similar to the one used in Mueser et al. (2007). This specification differs from our benchmark specification mainly in dropping detailed history information and rich personal information. Our detailed sensitivity analysis on history information showed that dropping this kind of information led only to small changes in the results as long as the requirement of the same elapsed unemployment duration before program start is not dropped (we do not drop this requirement here). But dropping rich personal information, such as information on family type and health status, leads to an overstatement of the treatment effects. In fact, part of the figures of the step 2 of sensitivity analysis 2 (appendix figs. E13/E14) look very similar to the figures of the sensitivity analysis involving the Mueser et al. (2007) specification for the propensity score. The remaining difference must be due to dropping even more information from the propensity

[33] Note, however, that the attitudinal variables in Dolton and Smith (2011) were directly elicited in a survey, while our variables are derived from information on program participation, receipt of unemployment benefits, and other information available to the caseworker. The population examined in Dolton and Smith (2011) also differs a lot from the one considered here.

score (mainly industry variables, a white collar dummy, and exact measurement of days in employment during the last 3 years) and to dropping detailed history information and rich personal information at the same time.[34]

Appendix figures E19/E20 and E21/E22 show the effects of omitting the propensity score altogether (steps 5 and 6). In fact, the results of these steps are remarkably similar to the results of the parsimonious Mueser et al. (2007) specification (appendix figs. E17/E18). In general, and particularly for women, estimated treatment effects are pulled upward considerably when omitting propensity score matching. This demonstrates that treated individuals are in general positively selected (because treatment effects are lower when personal characteristics are accounted for). The similarity between the estimates that omit the propensity scores altogether and the relatively parsimonious Mueser et al. (2007) specification reinforces the point that such a parsimonious specification may fail to provide unbiased treatment effect estimates. Furthermore, the similarity of the estimates from step 5 and step 6 underscores that this bias is not related to the omission of exact matching on employment histories.

As before, it is interesting to have a closer look at which of the variables representing rich personal information are significant determinants of the propensity score. In particular, it turns out that for both men and women, variables relating to family type, presence of children, and health or disability status significantly influence the probability of being treated (see appendix section B.1). In many cases, variables describing personality information also significantly enter the propensity score, although ignoring this information did not make much difference for our estimated treatment effects (see step 1 of the current sensitivity analysis). For example, information on whether there were any penalties or signs of lack of motivation mattered for both men and women in many of the cases considered. We also investigated to what extent treatment and comparison groups were unbalanced with respect to the personal information defined above (see appendix section B.2). It turns out that there were significant imbalances with respect to the second set of variables (children, family type, health, etc.) for men and women in the case of STT versus searching both in the raw data and after the imperfect matching of step 2. This is not, or to a much lesser extent, the case for the variables dropped in step 1 (past penalties, lack of motivation, etc.). For CFT versus searching, the number of significant imbalances before and after matching was low for both groups of variables.

[34] In line with our results, Lechner and Wunsch (2011) also find that the availability of a rich set of covariates, including, for example, information on health, job search behavior, and the timing of unemployment and program start, matters to account for the selection into treatment.

C. Sensitivity Analysis 3: Information on Other Programs

In comparison to the data used in Mueser et al. (2007) and in many other evaluation studies, our data are also more comprehensive in another important respect: they contain information on all possible programs a person might participate in. Hence, our third sensitivity analysis addresses the point to what extent the information on other programs matters. Specifically, we compare our benchmark specification to one that ignores information on programs other than the one in question. This mimics the situation in Mueser et al. (2007), which acknowledges that an unknown number of comparison group members may have participated in other programs. In a second step, we combine this with the Mueser et al. (2007) specification of the propensity score.

The results for the comparison with a situation where no information on other programs is available are shown in appendix figures E23/E24. The estimated effects are often pulled downward as in "CFT vs searching for men in stratum 2" and "STT vs searching for women in stratum 1." The reason for this seems to be that, if information on other programs is available, participants enrolling in the same stratum in these other programs are not in the comparison group. Including them in the comparison group pulls the estimated treatment effects upward to the extent that the participants in the other programs are locked into these programs and have low employment rates. When judging the value of this information, one should consider that the lack of information on other programs is likely to be less severe for countries with a relatively low program density (such as the United States) than for those continental European countries with a high program density.

The results for step 2 (in addition, parsimonious propensity score) are shown in appendix figures E25/E26. Here, the tendency to pull down estimated treatment effects by ignoring information about other programs is more than offset by the tendency to overstate treatment effects if the positive selection of training participants is not accounted for. Even if not completely generalizable to every institutional environment, our results demonstrate that limitations in data availability may have sizable effects on evaluation results.

D. Sensitivity Analysis 4: 12-Month Evaluation Window

The purpose of this section is to investigate how estimated treatment effects change when the width of the evaluation window is changed. In Section IV, we argued that a dynamic rather than a static evaluation setup was appropriate for our application. In the following, we define a 12-month evaluation stratum and compare its results to those obtained in our three-strata benchmark setup.[35] We expect two effects when going from our finer

[35] For example, Wunsch and Lechner (2008) use a 18-month stratum in their main analysis and provide results for a 12-month stratum in a sensitivity analysis.

stratification scheme to a single 12-month window. First, as pointed out by Fredriksson and Johansson (2008), excluding individuals from the comparison group who receive treatment within a time window of 12 months conditions on their future outcomes. Such future participants represent a negative selection because they would not receive treatment in the future if they had found employment before. Excluding them is therefore likely to pull down estimated treatment effects. Second, using a 12-month evaluation window does not account for potential interactions between elapsed unemployment duration and covariates, as only one propensity score is estimated for the whole sample rather than three different ones referring to the beginning of each stratum.

Appendix figure E27 presents the comparison between the results of a 12-month evaluation window and the results from our benchmark aggregated over the three strata used there.[36] Surprisingly, the two sets of results are rather close together. It turns out, however, that this is a coincidence and the result of two countervailing effects: excluding future participants pulls down estimated treatment effects, while using a pooled instead of a stratum-specific propensity score does the opposite. Appendix figures E28/E29 show the treatment effects obtained when we use our benchmark propensity scores but exclude comparison persons who receive training at a later point during the first 12 months of unemployment (step 2).[37] Appendix figures E30/E31 show the treatment effects obtained when we use the pooled propensity score in addition to excluding comparison persons who are treated within 12 months (step 3). As expected, excluding future participants pulls down the estimated treatment effects (step 2). The effect of using pooled rather than stratum-specific propensity scores is given by the difference between step 2 and step 3 (the two steps only differ in their use of the pooled vs. the stratum-specific propensity scores). It can be seen that using a pooled propensity score pulls up estimated treatment effects considerably.[38] Combining the effects of excluding future participants, using a pooled propensity score, and aggregating over strata yields the result of the 12-month evaluation window shown in appendix figure E27.

As a further sensitivity analysis related to a 12-month evaluation window, appendix figure E32 presents the results of a 12-month stratum but

In addition, they use the method of hypothetical program starts, which is investigated in the next section.

[36] For the aggregated benchmark, we use the individual treatment effects computed with the stratum-specific matching procedure and aggregate them across all participants in the three strata.

[37] We only exclude future participants when calculating the matching weights. The propensity scores are exactly as in the benchmark, i.e., they are estimated on samples that include participants in later strata.

[38] We found that the correlation between the pooled propensity score and the stratum-specific propensity scores decreased when going from earlier to later strata.

in which we omit matching on elapsed unemployment duration, history sequences, and calendar month of unemployment start. This drastically pulls down the estimated treatment effects, because the treated individuals are to a large extent matched to comparison persons who have already found employment. As expected, this effect is much larger than the one observed in step 5 of sensitivity analysis 1, in which we omit matching on elapsed unemployment duration within the smaller strata considered there. Finally, appendix figure E33 shows that this is exclusively due to the omission of elapsed unemployment duration and not due to the omission of history sequences and calendar month (the analysis still does not match on the latter, but it is already very close to the results of the original 12-month window given in step 1, in which we match on everything).

E. Sensitivity Analysis 5: Hypothetical Program Starts

In this sensitivity analysis, we compare our benchmark estimates (aggregated over the three strata) with the results from a 12-month evaluation window where we align comparison units to treated units according to their hypothetical program start (HPS). The method of hypothetical program starts is due to Lechner (1999) and is also used, for example, in Wunsch and Lechner (2008) and Lechner and Wunsch (2011). To construct the hypothetical program start dates for the comparison units, we regress the log program start dates of the treated persons on a set of covariates measured at the beginning of unemployment. Then we use the estimated coefficients plus a draw from the empirical distribution of the residuals to predict a start date for the nonparticipants. Nonparticipants whose predicted program start date lies after the end of their unemployment spell are dropped. The comparison group members are then matched based on the similarity of their hypothetical program start date with the start dates of the participants. In addition, the program start dates enter the propensity score model. Similar to Wunsch and Lechner (2008), we perform kernel matching with respect to the Mahalanobis distance of the estimated propensity score and the program start.

Our sensitivity analysis includes three steps. In step 1, we use a 12-month stratum in combination with the hypothetical program starts methodology where the program starts are included as dummies in the propensity score. In addition, we exactly match on the calendar month of unemployment start and on 4-year employment history sequences, as in our benchmark. In step 2, we omit the exact matching on calendar time and history sequences. Step 2 is closest to the scenario used in Wunsch and Lechner (2008). In a final step 3, we proceed as in step 2 but include the hypothetical program starts only linearly in the propensity score.

The results using the HPS methodology are shown in appendix section E.5 in the supplemental online appendix. The most conspicuous dif-

ferences to the benchmark relate to the more pronounced negative treatment effects. Qualitatively, they are much closer to those in Wunsch and Lechner (2008), who use the same data and study similar programs, than to our benchmark.[39] The comparison of steps 1–3 suggests that neither matching on 4-year history sequences and calendar time nor the way how the program start dates enter the propensity score make much difference. This is confirmed by further sensitivity analyses (not reported here) varying a number of further aspects of the HPS methodology. For example, we vary the random element in the construction of the start dates (by drawing from a different distribution or by resetting the seed). This made surprisingly little difference. We also impose exact matching on elapsed unemployment duration (represented by the hypothetical start dates in the case of comparison persons) or omit predicted start dates from the propensity score. Finally, we even change the time axis for the treatment effects from counting months since program start to counting months in unemployment. All this did not change the estimated treatment effects in a substantial way.

What explains yet the differences between the HPS results and our benchmark? The use of a 12-month stratum implies excluding future participants, who have disproportionately long unemployment spells, from the comparison group. This tends to pull down treatment effects (see step 2 of sensitivity analysis 4, appendix figs. E28/E29). As seen in step 3 of sensitivity analysis 4 (appendix figs. E30/E31), this effect happens to be offset by the use of a static propensity score, which tends to pull up estimated treatment effects. Our conjecture is that matching with respect to hypothetical programs starts allows one to account for dynamic selection effects in the otherwise static setup of the 12-month stratum (which is a good thing). In the HPS framework, treated individuals are matched with comparison group members who have a similar hypothetical program start date. As the program start date is predicted on the basis of observed characteristics, comparison group members will be similar in terms of their observed characteristics to individuals who receive treatment at a particular point in time. Thus, the downward bias from using a 12-month window is not alleviated by the use of a static propensity score. Actually, the HPS methodology yields even more negative effects than the results obtained for the 12-month window combined with matching on strata specific propensity scores (sensitivity analysis 4, step 2). This suggests that there is an additional negative shift that arises through matching on hypothetical start dates. Comparison units that are matched to early participants appear to be highly positively

[39] There are further differences that may lead to differences in estimated treatment effects. For example, Wunsch and Lechner (2008) pool men and women and consider only persons who receive unemployment benefits or unemployment assistance in the month of program start.

selected. We see this confirmed in our data as the majority of the controls with early hypothetical program starts exit unemployment within 3 months after their hypothetical program start. Thus, the HPS methodology seems to introduce a strong interaction between personal characteristics and unemployment experience.

F. Sensitivity Analysis 6: Future Program Participation

When estimating the effect of treatment versus searching, the question arises as to what extent our results are determined by the fact that the comparison group includes individuals who may participate in some program in the future (i.e., after the end of the stratum considered). Tables 6 and 7 provide some descriptive evidence on the incidence of later program participation in the comparison groups. Future participants make up between 11% and 20% of the raw comparison groups (see col. "Overall" in table 6). Their share increases further to 15%–27% when we weight the comparison observations with their matching weights (see table 7). For example, in the evaluation of "STT vs searching for men in stratum 1," future participants in the same or in another program represent 22.42% of the matched sample. Table 6 further demonstrates that the incidence of future participation declines over time because the shares in columns 3–5 remain approximately stable while the time intervals of elapsed unemployment lengthen from 3 to 12 months.

In the following decomposition analysis, we investigate the effects of sequentially excluding the outcomes of future participants from the pool of comparison units. This will shed light on how including or excluding future participation influences the results. In the first step, we proceed as in the benchmark but exclude the information on the outcomes of future participants from the end of their future program onward. In step 2 of our

Table 6
Importance of Later Program Participations

		Share of Future Participants (%)						
			By Time Period			By Program Type		
	Total Number of Controls	Overall	Month 4–6	Month 7–12	Year 2	STT	CFT	Other
Males:								
Stratum 1	29,351	17.89	5.47	6.44	5.97	6.35	2.33	9.21
Stratum 2	18,729	19.46		10.10	9.36	7.04	2.30	10.12
Stratum 3	11,348	15.45			15.45	5.78	1.39	8.28
Females:								
Stratum 1	18,409	17.92	5.69	7.07	5.16	6.98	2.90	8.04
Stratum 2	12,749	17.65		10.20	7.44	6.89	2.66	8.10
Stratum 3	8,795	10.79			10.79	4.38	1.55	4.86

NOTE.—Strata 1, 2, and 3 refer to programs starting in months 1–3, 4–6, and 7–12 of unemployment, respectively. All numbers refer to program versus searching comparisons.

Table 7
Percentage of Later Program Participation in Terms of Matching Weights

	Stratum 1	Stratum 2	Stratum 3
Males:			
STT versus searching	22.42	24.58	20.36
CFT versus searching	21.65	22.02	15.33
Females:			
STT versus searching	20.69	27.36	22.01
CFT versus searching	19.83	24.41	16.52

NOTE.—STT = short-term training; CFT = classroom further training. Strata 1, 2, and 3 refer to programs starting in months 1–3, 4–6, and 7–12 of unemployment, respectively. Shares are measured in percent of all potential comparison units (see col. 1 of table 6).

analysis, we exclude the information on their outcomes already from the start of their future program onward.[40] Finally, in step 3 of our analysis, we exclude future participants from the beginning of the stratum.[41] Results from this exercise are shown in appendix section E.6.

First, consider the differences between step 1 and the benchmark (in which future participants are fully included). In step 1, potentially positive posttreatment outcomes of future participants are excluded as they are deleted from the comparison group from the end of their future program onward. We would expect treatment effects in step 1 to be pulled upward compared with the benchmark estimates because future participants remain only in the comparison group so long as they have negative outcomes. However, in appendix figures E37/E38, treatment effects for step 1 lie clearly below those obtained in the benchmark case. Thus, reducing the weight of future participants in the control group results in a deterioration of treatment effects. This shows that potentially positive outcomes of future participants are unlikely to play any important role in our analysis. It rather implies that once we remove future participants, treated units are matched to a larger extent with those unemployed who quickly take up employment. In comparison, future participants represent a negative selection.

Now consider the difference between step 2 (censoring at the beginning of the future program) and step 1 (censoring at the end of the future program). Going from step 1 to step 2 means reducing further the weight of future participants in the comparison group. Consistent with the downward change from the benchmark to step 1, we now expect treatment effects to be pulled down even further. Indeed, treatment effects are more

[40] A similar approach has been suggested in Fredriksson and Johansson (2008) and de Luna and Johansson (2010).

[41] In this last step, we also estimate the propensity scores on samples that exclude the future participants.

negative in appendix figures E39/E40 compared with figures E37/E38. Finally, step 3 completely excludes future participants from the control group. Appendix figures E41/E42 show that treatment effects deteriorate again.

To sum up, participants enrolling later during unemployment constitute a disproportionately high share of the comparison group in terms of their matching weights. Removing them pulls estimated treatment effects down for two reasons. First, one systematically deletes persons known to have relatively bad outcomes because these people would not have received treatment later if they had already found employment again. Second, the next-best matches to treated units at a given point in time are those unemployed with relatively high chances to exit to employment in the near future.

G. Statistical Significance of the Sensitivity Analyses

To assess whether the differences between the benchmark and the sensitivity variants considered by us are statistically significant, we computed bootstrap standard errors for selected cases, jointly resampling the benchmark and the sensitivity estimates.[42] Results for the average differences between the benchmark and different sensitivity variants in year 1 and 2 after program start can be found in appendix section E.7. It turns out that the differences are in most cases statistically insignificant.[43] Notable exceptions are step 5 of sensitivity analysis 1 (no matching on elapsed unemployment duration) and step 3 of sensitivity analysis 6 (completely excluding future participants from the comparison group).

More important than whether or not differences between the benchmark and particular sensitivity variants are significant is whether the use of a particular sensitivity variant leads to conclusions that differ substantively from those of the benchmark. This may happen if treatment effects appear statistically significant (insignificant) in the sensitivity variant but insignificant (significant) in the benchmark. As we report in appendix section E.7, in a number of sensitivity cases the substantive conclusions would differ from the benchmark, for example, in step 5 of sensitivity analysis 1 (no matching on elapsed unemployment duration), in steps 2 and 4 of sensitivity analysis 2 (rich personal information), and in steps 2 and 3 of sensitivity analysis 6

[42] We implemented a semiparametric bootstrap, where we resampled persons from the combined samples of the benchmark and the sensitivity variants and repeated the matching analysis in every resample with a draw from the asymptotic distribution of the coefficients in the propensity score. We compared this procedure to that of the purely nonparametric bootstrap that we used for our benchmark estimates for all cases and found virtually no differences.

[43] This is also the general finding in Lechner and Wunsch (2011) in the part of the analysis that is comparable to ours.

(future program participation). Note that even if the difference between the benchmark and a given sensitivity variant is not significant in the vast majority of cases, it is still informative about the direction of possible biases and changes in the conclusions based on standard tests of significance of treatment effects.

VII. Conclusions

This article analyzes the employment and earnings effects of short-term training and classroom further training in West Germany. We employ a stratified propensity score matching approach to address dynamic selection into heterogeneous programs. We carefully assess to what extent various aspects of our empirical strategy, such as conditioning flexibly on employment and benefit histories, the availability of rich data, handling of later program participations, and further methodological choices, affect our estimates.

Our benchmark estimates suggest that the effectiveness of the different programs strongly depends on the gender of the participants and the timing of program participation during unemployment. We find statistically significant positive employment effects for male and female participants in short-term training and classroom further training who started their training not too early during their unemployment spell. Furthermore, we find significant earnings gains in the cases in which there are also positive employment effects with the earnings gains induced by classroom further training exceeding those induced by short-term training. Comparing the two program types, short-term training programs are (due to their much shorter durations) associated with much shorter lock-in periods and considerably lower total costs. There is some heterogeneity in the effects for different subpopulations, but we could not detect further effect heterogeneity regarding the personal characteristics of participants. As a caveat, our study does not account for potential general equilibrium effects (Calmfors 1994).

Our careful sensitivity analysis examines to what extent data features and methodological choices influence our impact estimates. We find that some data and specification issues can have a large effect. First, while conditioning on the exact unemployment experience in the current spell may have a large impact on the estimated treatment effects, flexibly conditioning on the employment history up to 7 years before the current unemployment spell makes little difference. Second, conditioning on information such as a lack of motivation, past benefit sanctions, or past program dropout does not affect estimated treatment effects. However, rich information on household type, number and age of dependent children, marital status, health, disability, and reasons why the last job ended is very important. Third, information on participation in programs other than the one under consideration is often important. Fourth, excluding future participants from the

control group reduces the estimated treatment effects. Fifth, using the alternative method of hypothetical program starts yields substantially more negative treatment effects compared with our benchmark estimates.

References

Abadie, Alberto, and Guido Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74:235–67.

——. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics* 29:1–11.

Barnow, Burt S. 1987. The impact of CETA programs on earnings: A review of the literature. *Journal of Human Resources* 22:157–93.

Bender, Stefan, Martin Biewen, Bernd Fitzenberger, Michael Lechner, Sonja Lischke, Ruth Miquel, Aderonke Osikominu, Tobias Wenzel, and Conny Wunsch. 2004. Die Beschäftigungswirkung der FbW-Maßnahmen 2000–2002 auf individueller Ebene: Eine Evaluation auf Basis der prozessproduzierten Daten des IAB–Vorläufiger Zwischenbericht Oktober 2004. Department of Economics, Goethe University Frankfurt, and SIAW, University of St. Gallen.

Bender, Stefan, Martin Biewen, Bernd Fitzenberger, Michael Lechner, Ruth Miquel, Aderonke Osikominu, Marie Waller, and Conny Wunsch. 2005. Die Beschäftigungswirkung der FbW-Maßnahmen 2000–2002 auf individueller Ebene: Eine Evaluation auf Basis der prozessproduzierten Daten des IAB–Zwischenbericht Oktober 2005. Department of Economics, Goethe University Frankfurt, and SIAW, University of St. Gallen.

Bender, Stefan, Annette Haas, and Christoph Klose. 2000. The IAB Employment Subsample, 1975–1995. *Schmollers Jahrbuch* 120:649–62.

Bergemann, Annette, Bernd Fitzenberger, and Stefan Speckesser. 2009. Evaluating the dynamic employment effects of training programs in East Germany using conditional differences-in-differences. *Journal of Applied Econometrics* 24:797–823.

Bergemann, Annette, and Gerard J. van den Berg. 2008. Active labor market policy effects for women in Europe: A survey. *Annales d'Economie et de Statistique* 91–92:385–408.

Biewen, Martin, Bernd Fitzenberger, Aderonke Osikominu, and Marie Waller. 2007. Which program for whom: Evidence on the comparative effectiveness of public sponsored training programs in Germany. IZA Discussion Paper no. 2885, Institute for the Study of Labor, Bonn.

Blaschke, Dieter, and Hans-Eberhard Plath. 2000. Möglichkeiten und Grenzen des Erkenntnisgewinns durch Evaluation aktiver Arbeitsmarktpolitik. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung* (*MittAB*) 33:462–82.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 1997. The benefits and

costs of JTPA Title II-A programs. *Journal of Human Resources* 32:549–76.
Blundell, Richard, Monica Costa Dias, Costas Meghir, and John van Reenen. 2004. Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association* 2:569–606.
Bundesagentur für Arbeit. 2001, 2002a, 2003a, 2004a, 2005a. *Daten zu den Eingliederungsbilanzen 2000, 2001, 2002, 2003, 2004*, Nürnberg (various issues).
——. 2002b, 2003b, 2004b, 2005b. *Arbeitsmarkt, 2001, 2002, 2003, 2004*. Nürnberg (various issues).
——. 2005c. *Geschäftsbericht, 2004*. Nürnberg.
Busso, Matias, John DiNardo, and Justin McCrary. 2009. Finite sample properties of semiparametric estimators of average treatment effects. Unpublished manuscript, Department of Economics, University of California, Berkeley.
——. 2011. New evidence on the finite sample properties of propensity score matching and reweighting estimators. Unpublished manuscript, Department of Economics, University of California, Berkeley.
Calmfors, Lars. 1994. Active labour market policy and unemployment: A framework for the analysis of crucial design features. *OECD Economic Studies* 22:7–47.
Carcillo, Stéphane, and David Grubb. 2006. From inactivity to work: The role of active labour market policies. OECD Social Employment and Migration Working Papers no. 36, Organization for Economic Cooperation and Development, Paris.
Card, David, Jochen Kluve, and Andrea Weber. 2010. Active labour market policy evaluations: A meta-analysis. *Economic Journal* 120:F452–F477.
Card, David, and Daniel Sullivan. 1988. Measuring the effect of subsidized training programs on movements in and out of employment. *Econometrica* 56:497–530.
Crépon, Bruno, Muriel Dejemeppe, and Marc Gurgand. 2005. Counseling the unemployed: Does it lower unemployment duration and recurrence? IZA Discussion Paper no. 1796, Institute for the Study of Labor, Bonn.
Crépon, Bruno, Marc Ferracci, Grégory Jolivet, and Gerard J. van den Berg. 2009. Active labor market policy effects in a dynamic setting. *Journal of the European Economic Association* 7:595–605.
de Luna, Xavier, and Per Johansson. 2010. Nonparametric inference for the effect of a treatment on survival times with application in the health and social sciences. *Journal of Statistical Planning and Inference* 140:2122–37.
Dolton, Peter, and Jeffrey Smith. 2011. The impact of the UK New Deal for Lone Parents on benefit receipt. IZA Discussion Paper no. 5491, Institute for the Study of Labor, Bonn.
Dorsett, Richard. 2006. The New Deal for Young People: Effect on the labour market status of young men. *Labour Economics* 13:405–22.

Dyke, Andrew, Carolyn J. Heinrich, Peter R. Mueser, Kenneth R. Troske, and Kyung-Seong Jeon. 2006. The effects of welfare-to-work program activities on labor market outcomes. *Journal of Labor Economics* 24: 567–607.

Fitzenberger, Bernd, Olga Orlanski, Aderonke Osikominu, and Marie Paul. 2013. Déjà vu? Short-term training in Germany, 1980–1992 and 2000–2003. *Empirical Economics* 44:289–328.

Fitzenberger, Bernd, Aderonke Osikominu, and Robert Völter. 2006. Imputation rules to improve the education variable in the IAB Employment Subsample. *Journal of Applied Social Science Studies* 126: 405–36.

———. 2008. Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany. *Annales d'Economie et de Statistique* 91–92:321–55.

Fitzenberger, Bernd, and Stefan Speckesser. 2007. Employment effects of the provision of specific professional skills and techniques in Germany. *Empirical Economics* 32:529–73.

Fitzenberger, Bernd, and Robert Völter. 2007. Long run effects of training programs for the unemployed in East Germany. *Labour Economics* 14: 730–55.

Fougère, Denis, Jacqueline Pradel, and Muriel Roger. 2005. Does job search assistance affect search effort and outcomes? IZA Discussion Paper no. 1825, Institute for the Study of Labor, Bonn.

Fredriksson, Peter, and Per Johansson. 2008. Dynamic treatment assignment: The consequences for evaluations using observational data. *Journal of Business and Economic Statistics* 26:435–45.

Frölich, Markus. 2004. Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics* 86:77–90.

——. 2008. Statistical treatment choice: An application to active labour market programmes. *Journal of American Statistical Association* 103:547–58.

Galdo, Jose, Jeffrey Smith, and Dan A. Black. 2008. Bandwidth selection and the estimation of treatment effects with unbalanced data. *Annales d'Economie et Statistique* 91–92:189–216.

Gerfin, Michael, and Michael Lechner. 2002. Microeconometric evaluation of the active labor market policy in Switzerland. *Economic Journal* 112:854–93.

Hardoy, Inés. 2005. Impact of multiple labour market programmes on multiple outcomes: The case of Norwegian youth programmes. *LABOUR* 19:425–67.

Heckman, James J., and V. Joseph Hotz. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84:862–74.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 65:1017–98.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64:605–54.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. The economics and econometrics of active labor market programs. In *Handbook of labor economics*, vol. 3A, ed. Orley Ashenfelter and David Card, 1865–2097. Amsterdam: North Holland.

Heckman, James J., and Jeffrey Smith. 1999. The pre-programme earnings dip and the determinants of participation in a social programme: Implications for simple programme evaluation strategies. *Economic Journal* 109:313–48.

Heinrich, Carolyn J., Peter R. Mueser, Kenneth Troske, Kyung-Seong Jeon, and Daver C. Kahvecioglu. 2010. New estimates of public employment and training program net impacts: A nonexperimental evaluation of the Workforce Investment Act program. Working Paper no. 1003, Department of Economics, University of Missouri.

Hotz, V. Joseph, Guido Imbens, and Jacob A. Klerman. 2006. Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. *Journal of Labor Economics* 24:521–66.

Huber, Martin, Michael Lechner, and Conny Wunsch. 2010. How to control for many covariates? Reliable estimators based on the propensity score. IZA Discussion Paper no. 5268, Institute for the Study of Labor, Bonn.

Hui, Shek-wai, and Jeffrey Smith. 2003a. Issues in the design of Canada's Adult Education and Training Survey. Research Report, Statistics Canada.

———2003b. The labour market impacts of adult education and training in Canada. Research report, Statistics Canada.

Hujer, Reinhard, Stephan L. Thomsen, and Christopher Zeiss. 2006. The effects of short-term training measures on the individual unemployment duration in West Germany. ZEW Discussion Paper no. 06-065, Centre for European Economic Research, Mannheim.

Human Resources and Skills Development Canada (HRSDC). 2004. Summative evaluation of employment benefits and support measures under the terms of the Canada/British Columbia Labour Market Development Agreement. Final report. Document no. SP-AH-666-04-04E.

Imbens, Guido. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87:706–10.

Jacobebbinghaus, Peter, and Stefan Seth. 2007. The German Integrated Employment Biographies Sample (IEBS). *Journal of Applied Social Science Studies* 127:335–42.

Jespersen, Svend T., Jakob R. Munch, and Lars Skipper. 2008. Costs and benefits of Danish active labour market programmes. *Labour Economics* 15:859–84.

Kluve, Jochen, and Claus M. Schmidt. 2002. Can training and employment subsidies combat European unemployment? *Economic Policy* 35:409–48.

Kurtz, Beate. 2003. Trainingsmaßnahmen: Was verbirgt sich dahinter? IAB Werkstattbericht Nr. 8, Institute for Employment Research, Nürnberg.

Larsson, Laura. 2003. Evaluation of Swedish youth labor market programs. *Journal of Human Resources* 38:891–927.

Lechner, Michael. 1999. Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics* 17:74–90.

——. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of active labor market politics in Europe*, ed. Michael Lechner and Friedhelm Pfeiffer, 45–58. Heidelberg: Physica-Verlag.

——. 2002. Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics* 84:205–20.

Lechner, Michael, Ruth Miquel, and Conny Wunsch. 2007. The curse and blessing of training the unemployed in a changing economy: The case of East Germany after unification. *German Economic Review* 8:468–507.

——. 2011. Long-run effects of public sector sponsored training in West Germany. *Journal of the European Economic Association* 9:742–84.

Lechner, Michael, and Stephan Wiehler. 2011. Kids or courses? Gender differences in the effects of active labor market policies. *Journal of Population Economics* 24:783–812.

Lechner, Michael, and Conny Wunsch. 2009. Active labour market policy in East Germany: Waiting for the economy to take off. *Economics of Transition* 4:661–702.

——. 2011. Sensitivity of matching-based program evaluations to the availability of control variables. IZA Discussion Paper no. 5553, Institute for the Study of Labor, Bonn.

Lee, Wang-Sheng. 2013. Propensity score matching and variations on the balancing test. *Empirical Economics* 44:47–80.

Martin, John P. 2000. What works among active labour market policies: Evidence from OECD countries' experiences. *OECD Economic Studies* no. 30, 79–113.

Martin, John P., and David Grubb. 2001. What works for whom: A review of OECD countries' experiences with active labour market policies. *Swedish Economic Policy Review* 8:9–56.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky. 2007. Using state administrative data to measure program performance. *Review of Economics and Statistics* 89:761–83.

OECD (Organization for Economic Cooperation and Development). 2005. Labour market programmes and activation strategies: Evaluating the impacts. In *OECD Employment Outlook*, chap. 4. Paris: OECD.

Osikominu, Aderonke. 2013. Quick job entry or long-term human capital development? The dynamic effects of alternative training schemes. *Review of Economic Studies* 80:313–42.

Park, Norm, Bob Power, W. Craig Riddell, and Ging Wong. 1996. An assessment of the impact of government-sponsored training. *Canadian Journal of Economics* 29:S93–S98.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Roy, Andrew D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3:135–46.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.

Schneider, Hilmar, Karl Brenke, Lutz Kaiser, Jakob Steinwedel, Birgit Jesske, and Arne Uhlendorff. 2006. Evaluation der Maßnahmen zur Umsetzung der Vorschläge der Hartz-Kommission; Modul 1b: Förderung beruflicher Weiterbildung und Transferleistungen. IZA Research Report no. 7, Institute for the Study of Labor, Bonn.

Sianesi, Barbara. 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. *Review of Economics and Statistics* 86: 133–55.

——. 2008. Differential effects of Swedish active labour market programmes for unemployed adults in the 1990s. *Labour Economics* 15: 370–99.

Splawa-Neyman, Jerzy (edited and translated by Dorota M. Dabrowska and Terrence P. Speed). 1923/1990. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science* 5, no. 4:465–72.

van den Berg, Gerard J., and Bas van der Klaauw. 2006. Counseling and monitoring of unemployed workers: Theory and evidence from a controlled social experiment. *International Economic Review* 47:895–936.

Waller, Marie. 2008. On the importance of correcting reported end dates of labor market programs. *Journal of Applied Social Science Studies* 128: 213–36.

Weber, Andrea, and Helmut Hofer. 2004. Employment effects of early interventions on job search programs. IZA Discussion Paper no. 1076, Institute for the Study of Labor, Bonn.

Wunsch, Conny, and Michael Lechner. 2008. What did all the money do? On the general ineffectiveness of recent West German labour market programmes. *Kyklos* 61:134–74.