The Employment Effects of Job Creation Schemes in Germany - A Microeconometric Evaluation*

Marco Caliendo[†], Reinhard Hujer[‡] and Stephan L. Thomsen[§]

[†]*DIW, Berlin and IZA, Bonn*

[‡]*J.W.Goethe-University, Frankfurt, IZA, Bonn, ZEW, Mannheim*

[§]*J.W.Goethe-University, Frankfurt*

Abstract

In this paper we evaluate the employment effects of job creation schemes on the participating individuals in Germany. Job creation schemes are a major element of active labour market policy in Germany and are targeted at long-term unemployed and other hard-to-place individuals. Access to very informative administrative data of the Federal Employment Agency justifies the application of a matching estimator and allows to account for individual (group-specific) and regional effect heterogeneity. We extend previous studies in four directions. First, we are able to evaluate the effects on regular (unsubsidised) employment. Second, we observe the outcome of participants and non-participants for nearly three years after programme start and can therefore analyse mid- and long-term effects. Third, we test the sensitivity of the results with respect to various decisions which have to be made during implementation of the matching estimator, e.g. choosing the matching algorithm or estimating the propensity score. Finally, we check if a possible occurrence of 'unobserved heterogeneity' distorts our interpretation. The overall results are rather discouraging, since the employment effects are negative or insignificant for most of the analysed groups. One notable exception are long-term unemployed individuals who benefit from participation. Hence, one policy implication is to address programmes to this problem group more tightly.

Keywords: Evaluation, Matching, Sensitivity Analysis, Job Creation Schemes, Long-term Unemployed
JEL Classification: J68, H43, C13
Word Count: approx. 9.800

---

*The authors thank Barbara Sianesi for valuable comments. The paper has also benefited from the discussion at the SOLE/EALE world conference in San Francisco (2005). The usual disclaimer applies. This paper emerged within the research project 'Effects of Job Creation and Structural Adjustment Schemes' financed by the Institute for Employment Research (IAB). Corresponding author: Marco Caliendo, DIW Berlin, Dep. of Public Economics, Königin-Luise-Str. 5, 14195 Berlin, Germany.

[†]Marco Caliendo is Senior Research Associate at the German Institute for Economic Research (DIW) in Berlin and Research Fellow of the IZA, Bonn, e-mail: mcaliendo@diw.de.

[‡]Reinhard Hujer is Professor of Statistics and Econometrics at the J.W.Goethe-University of Frankfurt, and Research Fellow of the IZA, Bonn and the ZEW, Mannheim, e-mail: hujer@wiwi.uni-frankfurt.de.

[§]Stephan L. Thomsen is Research Assistant at the Institute of Statistics and Econometrics, J.W.Goethe-University of Frankfurt, e-mail: sthomsen@wiwi.uni-frankfurt.de.

1 Introduction

The German labour market is plagued by persistently high unemployment in combination with a clearly separated situation on the labour markets as the unemployment rates of 9.3% in West and 20.1% in East Germany reflect. The Federal Employment Agency (FEA) spends substantial on active labour market policies (ALMP) to overcome this unemployment unemployment problem (West: 12.3 bn Euro, East: 8.9 bn Euro). The main goal of ALMP is the permanent integration of unemployed persons into regular employment. Since 1998 the legal basis for ALMP is the Social Code III ('Sozialgesetzbuch III') replacing the Work Support Act from 1969. Within that reform new instruments were introduced, competencies were decentralised and a more flexible allocation of funds has been made possible. Maybe the most important change from an evaluator's point of view was the legal anchoring of a mandatory output evaluation for all ALMP measures.[1]

One major element of ALMP in Germany over the last years have been job creation schemes (JCS). Their importance is currently decreasing as they have often been criticised for the lack of explicit qualificational elements and 'stigma effects'.[2] Though, it can also be argued that they are a reasonable opportunity for individuals who are not able to re-integrate into the first labour market themselves or who do not fit the criteria for other programmes, e.g. long-term unemployed or other hard-to-place individuals. The evaluation of JCS in Germany has been impossible for a long time, since datasets have either not been available or been to small to draw policy relevant conclusions. However, with the introduction of the Social Code III things have changed and give us access to a very rich administrative dataset containing more than 11,000 participants in JCS and a comparison group of nearly 220,000 non-participants. We use this data to answer the question, if JCS enhance the employment chances of participating individuals. The extensive set of available individual characteristics in combination with information on the regional labour market situation makes the application of a matching estimator

[1] The reform process on the German labour market is still ongoing. More reforms are implemented gradually (see 'Modern Services on the Labour Market', Bundesministerium für Wirtschaft und Arbeit (2003)). Since we focus in our empirical analysis on the time period 2000-2002, we are not going to discuss the current reforms here.

[2] If the programme is targeted at people with 'disadvantages', there is always a risk that a possible employer takes participation in such schemes as a negative signal concerning the expected productivity or motivation.

possible. Additionally, the large number of participants allows to account for several sources of effect heterogeneity.

The importance of effect heterogeneity for the evaluation of JCS in Germany has been well documented in Hujer, Caliendo, and Thomsen (2004). Basically, there are two shortcomings to that study. The first one refers to an unsatisfying outcome variable, which allows only to monitor if the individual is registered unemployed or not and does not allow to draw conclusions about the re-integration success into regular (unsubsidised) employment. A second restriction relates to the relatively short observation period after programme start, namely two years. This paper extends previous analyses in four important directions. First, we are able to evaluate the re-integration effects of JCS into regular (unsubsidised) employment. Second, we can monitor the employment status of participants and non-participants nearly three years after programme start. Third, we test the sensitivity of the results to various decisions which have to be made whilst implementing the matching estimator, like the choice of the matching algorithm or the estimation of the propensity score. Finally, we check if a possible occurrence of 'unobserved heterogeneity' or 'hidden bias' distorts interpretation of our results.

The focus of our analysis will be the identification of individual (group-specific) and regional effect heterogeneity. To do so, we separate the analysis by several characteristics and carry out the matching analysis on sub-populations.[3] Men and women in West and East Germany will be the 'main groups' of our analysis. In addition, we estimate effects for eleven 'sub-groups' defined by age and unemployment duration as well as by specific characteristics indicating disadvantages on the labour market like the lack of professional training or the existence of placement restrictions due to health problems.

The remainder of this paper is organised as follows. In the following section we describe the institutional background of JCS in Germany, introduce the dataset used and present some descriptive statistics. Section 3 explains the general framework for microeconometric evaluation analysis and section 4 deals with the empirical implementation of the matching estimator. In particular we discuss

[3] There are basically two ways to put greater emphasis on specific variables. One can either find variables in the comparison group who are identical with respect to these variables (see e.g. Puhani (1998) or carrying out matching on sub-populations (see e.g. Heckman, Ichimura, and Todd (1997) or Heckman, Ichimura, Smith, and Todd (1998)).

the justification of the matching estimator, the estimation of the propensity scores (section 4.1) and the choice of the proper matching algorithm (section 4.2) for our situation. Section 4.3 deals with common support issues, whereas section 4.4 presents some quality indicators for the chosen matching algorithm. In section 5 we present the results for the main and sub-groups. Additionally, we also test the sensitivity of our estimates with respect to unobserved heterogeneity. The final section concludes and gives some policy recommendations.

2 Institutional Background, Dataset and Selected Descriptives

2.1 Institutional Background

JCS have been one major element of ALMP in Germany. They can be supported if they provide the last chance to stabilise and qualify unemployed individuals for later re-integration into regular employment.[4] JCS should support activities, which are additional in nature and for the collective good. Additional in nature means, that the activities could not be executed without the subsidy. To prevent substitution effects and windfall gains, measures with a predominantly commercial purpose have been excluded up to January 2002, when the legal requirements have been relaxed by a law amendment ('Job-AQTIV-Gesetz'). The majority of activities is conducted in the public and non-commercial sector. Since JCS are co-financed measures, support is obtained as a wage subsidy to the employer, covering 30% to 75% of the costs, which is usually paid for twelve months. Exceptions can be made for a higher subsidy-quota (up to 100%) and a longer duration (up to 24 months) for persons with strong labour market disadvantages or projects of high priority (see below). Before January 2002, potential participants had to be long-term unemployed (for more than one year) or unemployed for at least six of the last twelve months before participation. Furthermore, they had to fulfil the requirements for entitlement of unemployment compensation. In addition, the local placement officers were allowed to place up to five percent of the allocated individuals who did not meet these conditions (Five-Percent-Quota). Further exceptions are made for young unemployed (under 25 years) without

[4] The legal basis of JCS is §§260-271, 416 Social Code III.

professional training, short-term unemployed (with at least three months of unemployment) placed as tutors, and disabled who could be stabilised or qualified.[5] Participants are allowed to do a practical training up to 40% of the time and a vocational training of up to 20%, together no more than 50% of the programme duration. Priority should be given to projects which enhance the chances for permanent jobs, support structural improvement in social or environmental services or aim at the integration of extremely hard-to-place individuals. Participation in JCS results from placement by the local labour office. Unemployed individuals, who fulfil the eligibility criteria, can be offered a specific occupation. The responsible caseworker can cancel a running programme at any time, if the participant can be placed into regular employment. If an unemployed person rejects the offer of a JCS or if a participant denies a career counselling by the placement officer, the labour office can stop unemployment benefits at the first time for up to twelve weeks. If rejections to offers repeat, persons may lose their entitlement for unemployment benefits.

2.2 Dataset

The empirical analysis is based on a dataset merged from several administrative sources of the FEA.[6] It contains information on all participants who have started a JCS in February 2000. The sample of non-participants was drawn from the Job-Seekers-Data-Base in January 2000. Hence, it consists of individuals who have been eligible for participation, but have not entered those schemes in February 2000. The dataset includes numerous attributes to describe the labour market situation of the individuals. The information can be categorised into four classes: socio-demographic information, like age, gender and marital status, qualification details (work experience, professional training, etc.), a short labour market history (durations of last employment and unemployment spells) and particular programme information (programme sector, programme duration). Since all information originates from the same sources for participants and non-participants, the dataset provides a good basis for

[5] With the 2002 amendment, all unemployed individuals can enter a JCS independently of the preceding unemployment duration, but with restriction that JCS is the only opportunity for occupation. In addition, the Five-Percent-Quota has been augmented to ten percent.

[6] See Caliendo, Hujer, and Thomsen (2005) for more details on the data sources.

the construction of a valid comparison group. The dataset is completed by a FEA characterisation of the regional labour market situation in the labour office districts (Blien *et al.*(2004)).[7] The outcome variable is taken from the Employment Statistics Register, which contains information on all regularly employed persons and participants of several ALMP programmes. In the analysis we treat only regular (unsubsidised) employment as a success, i.e. subsidised employment or participation in other ALMP programmes are defined as a failure. Although this definition is in conflict with the institutional setting, it reflects the economic point of view to measure the integration ability of JCS into regular (unsubsidised) employment.[8] We observe the labour market outcome for all individuals until December 2002 and the final sample consists of 11,151 participants and 219,622 non-participants.[9] As previous empirical findings have shown that the effects of JCS differ with respect to region and gender (see e.g. Hujer, Caliendo, and Thomsen (2004)), we separate our analysis by these aspects, i.e. we estimate the effects separately for men and women in West and East Germany, which are the four 'main groups' of our analysis. Table 1 shows that the largest groups are women (5,035) and men (2,924) in East Germany. In West Germany, 2,140 men and 1,052 women have started a JCS in February 2000.

Insert table 1 about here

Due to the large number of observations in our sample, we are also able to analyse the programme effects for specific problem-groups of the labour market. We distinguish three age categories (younger than 25 years, between 25 and 50 years, and older than 50 years), three unemployment durations (up to 13 weeks, between 13 and 52 weeks, and for more than 52 weeks), persons without work experience, without professional training and the counterfactual group of persons with a high educational degree (college and university graduates). In addition, we analyse the effects for rehabilitation attendants, and

[7] This classification assigns the 181 labour office districts in Germany to twelve comparable clusters, based on characteristics like the underemployment quota, population density and the number of vacancies.

[8] Only the first programme participation is evaluated, any participation in later programmes is viewed as an outcome of the first treatment and is defined as a failure.

[9] Information on participants in Berlin is excluded, since the special situation of the labour market in the capital city would require a separate evaluation of the effects. However, the small number of participants aggravates the interpretation of the results.

for individuals for whom the caseworkers have noted placement restrictions due to health problems. In total we get eleven 'sub-groups' for whom the effects will be estimated separately in both regions and for both gender. Table 1 contains the observations in these groups, differentiated by participation status. What can be seen as most important is that nearly all groups contain a reasonable number of participants, allowing a proper estimation and interpretation of the effects.[10]

The figures from table 1 have shown that the majority of persons participate in East Germany. Looking at the individual characteristics measured at programme begin in February 2000 shows regional differences, too. To save space we refrain from doing so here, but we want to mention one important finding. The descriptives show, that the allocation of individuals to JCS is not target-oriented. The results also indicate a different purpose of JCS in both regions, where JCS are more target-oriented (e.g. for young unemployed without professional training) in West-Germany and are also used to relieve the tense situation on the labour market in East Germany.[11]

3 The Estimation of Treatment Effects with Matching Estimators

The Potential Outcome Framework and Selection Bias Since we work with non-experimental data, we have to deal with some identifying issues. As we consider only one specific programme compared to non-participation, we can use the potential outcome framework with two potential outcomes Y^1 (individual receives treatment) and Y^0 (individual does not receive treatment).[12] The actually observed outcome for any individual i can be written as: $Y_i = Y_i^1 \cdot D_i + (1 - D_i) \cdot Y_i^0$, where $D \in \{0, 1\}$ is a binary treatment indicator. The treatment effect for each individual i is then defined as the difference between her potential outcomes $\Delta_i = Y_i^1 - Y_i^0$. Since there will be never an opportunity to estimate individual effects with confidence, we have to concentrate on population averages of gains from treatment. The most prominent evaluation parameter is the so called average treatment effect on the treated (ATT), which focuses explicitly on the effects on those for whom the programme is

[10] Groups with less than 100 observations are excluded from the analysis.

[11] Tables with selected descriptives for the main groups are available on request by the authors.

[12] The potential outcome framework is variously attributed to Fisher (1935), Neyman (1935), Roy (1951), Quandt (1972, 1988) or Rubin (1974), but most often it is just called the Roy-Rubin-model (RRM).

actually intended. It is given by:

$$ATT = E(\Delta \mid D = 1) = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1). \quad (1)$$

Given equation (1) the problem of selection bias is straightforward to see, since the second term on the right hand side is unobservable. If the condition $E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0)$ holds, we can use the non-participants as an adequate comparison group. However, with non-experimental data it will usually not hold. Consequently, estimating the ATT by the difference in the subpopulation means of participants $E(Y^1 \mid D = 1)$ and non-participants $E(Y^0 \mid D = 0)$ will lead to a selection bias, since participants and non-participants are selected groups that would have different outcomes even in absence of the programme. This bias may come from observable factors like age or skill differences or unobservable factors like motivation. For both cases different estimation strategies are available.[13] If one is willing to assume that selection occurs on observed characteristics only, the matching estimator is an appealing choice. Its basic idea is to search from a large group of non-participants those individuals who are similar to the treated group in all relevant (observable) characteristics.[14]

How Does Matching Solve the Evaluation Problem? Matching is based on the conditional independence (or unconfoundedness) assumption (CIA) which states, that conditional on some covariate X, the outcome Y is independent of D. Since matching on X can become hazardous when X is of high dimension ('curse of dimensionality'), Rosenbaum and Rubin (1983) suggest the use of balancing scores $b(X)$. These are functions of the relevant observed covariates X such that the conditional distribution of X given $b(X)$ is independent of the assignment to treatment, that is $X \amalg D|b(X)$. For participants and non-participants with the same balancing score, the distributions of the covariates X are the same, i.e. they are balanced across the groups. The propensity score $P(X)$, i.e. the probability of participating in a programme is one possible balancing score. It summarises the information of the observed covariates X into a single index function. The propensity score can be seen as the coarsest

[13] See Heckman, LaLonde, and Smith (1999), Angrist and Krueger (1999) or Blundell and Costa Dias (2002) for overviews.
[14] See Imbens (2004) or Smith and Todd (2005) for recent overviews regarding matching methods.

balancing score whereas X is the finest (Rosenbaum and Rubin, 1983). The authors also show that if treatment assignment is strongly ignorable given X, it is also strongly ignorable given any balancing score. Hence, it is sufficient to assume that (in the notation of Dawid (1979)):

Assumption 1 *Unconfoundedness for Controls given the Propensity Score:*

$$Y^0 \amalg D|P(X),$$

where $\amalg$ denotes independence. If assumption 1 is fulfilled, the non-participant outcomes have, conditional on $P(X)$, the same distribution that participants would have experienced if they had not participated in the programme (Heckman, Ichimura, and Todd, 1997). Similar to randomisation in a classical experiment, matching balances the distributions of all relevant pre-treatment characteristics X in the treatment and comparison group, and thus achieves independence between the potential outcomes and the assignment to treatment. Hence, if the mean exists,

$$E(Y^0 \mid X, D = 1) = E(Y^0 \mid X, D = 0) = E(Y^0 \mid X)$$

and the missing counterfactual mean can be constructed from the outcomes of non-participants. In order for both sides of the equations to be well defined simultaneously for all X, it is usually additionally assumed that

Assumption 2 *Weak Overlap:*

$$Pr(D = 1 \mid X) < 1,$$

for all X. This implies that the support of X is equal in both groups, i.e. $S = \text{Support}(X|D = 1) = \text{Support}(X|D = 0)$. These assumptions are sufficient for identification of (1), because the moments of the distribution of Y^1 for the treated are directly estimable.[15] The method of matching can also be used to estimate the ATT at some points $X = x$, where x is a particular realisation of X:

$$ATT(X = x) = E(\Delta \mid X = x, D = 1) = E(Y^1 \mid X = x, D = 1) - E(Y^0 \mid X = x, D = 1). \qquad (2)$$

This parameter measures the mean treatment effect for persons who were randomly drawn from the population of the treated given a specific realisation of certain characteristics X. This is of particular

[15] To identify the average treatment effect (ATE), additional assumptions are required.

interest for us, since we want to estimate the effects for sub-groups, like long-term unemployed persons or individuals without work experience.

4 Implementation of the Matching Estimator

After having decided to use matching estimators for evaluation purposes, the researcher is confronted with several questions regarding the implementation of these estimators.[16] Every evaluation task requires a careful consideration of the available choices for the given situation. Hence, we will discuss the implementation and justification of the matching estimator in our context in the next subsections. We start with the plausibility of the CIA in our context. Since the number of covariates in the data makes the use of covariate matching unfeasible, we rely on propensity score matching. To do so, we have to consider the correct model and the choice of relevant variables for the participation probability. Following that, we choose one matching algorithm to be used in the further analysis. Subsections 4.3 and 4.4 will be concerned with common support and matching quality issues.

4.1 Plausibility of CIA and Propensity Score Estimation

Before estimating the propensity scores, we have to consider the plausibility of the CIA in our context briefly. Our dataset contains a rich set of variables, including socio-demographic variables, information about the qualificational background and the labour market history of individuals. The latter point is most important since previous empirical studies have emphasised the importance of the labour market history (see Heckman, LaLonde, and Smith (1999)). Finally, the situation on the regional labour market is accounted for by using the FEA-clusters as described in Blien *et al.*(2004). Given this informative dataset we argue henceforth that the CIA holds. We test the sensitivity of our estimates to this assumption using a bounding analysis in section 5.

Estimation of the propensity scores requires two choices to be made: First, the choice of an adequate model and second, the selection of variables to be included in this model. Little advice is available regarding which functional form to use (see e.g. Smith (1997)). In principle any discrete choice model

[16] Caliendo and Kopeinig (2005) provide an extensive overview of the issues arising when implementing matching estimators.

can be used. For the binary treatment case logit and probit models usually yield similar results making the choice not too critical. Since the logit distribution has more density mass in the bounds that reflects our situation better, it will be used for estimation.

More advice is available regarding the inclusion (or exclusion) of covariates in the propensity score model. Only variables that simultaneously influence the participation decision and the outcome variable should be included. Economic theory, a sound knowledge of previous research and information about the institutional settings should guide the researcher in building up the model (see e.g. Smith and Todd (2005) or Sianesi (2004)). Furthermore, it should be clear that only variables that are unaffected by participation (or the anticipation of it) should be included in the model. Economic theory gives some guidance on which variables to choose. The accumulated evidence in the evaluation literature points out that the labour market history of individuals and the regional labour market environment are crucial variables to be included in the estimation (Heckman, LaLonde, and Smith, 1999). In cases of uncertainty of the proper specification, the question might arise if it is better to include too many rather than too few variables. The literature is ambiguous about this point. Whereas Bryson, Dorsett, and Purdon (2002) argue that over-parameterised models should be avoided, Rubin and Thomas (1996) recommend against 'trimming' models in the name of parsimony. But clearly, there are also some formal (statistical) tests which can be used. Table 2 contains the results for the hit-rate and the pseudo-R^2. With the 'hit or miss'-method or prediction rate metric (suggested by Heckman, Ichimura, Smith, and Todd (1998) and Heckman and Smith (1999)), variables are chosen to maximise the within-sample correct prediction rates, assuming that the costs for the misclassification are equal for both groups (Heckman, Ichimura, and Todd, 1997).[17]

Insert table 2 about here

Both statistics have been estimated for several specifications of the model. We have started with base specifications, containing only variables of one of the four above mentioned categories. Using this

[17] See e.g. Breiman, Friedman, Olsen, and Stone (1984) for theory and Heckman, Ichimura, Smith, and Todd (1998) or Smith and Todd (2005) for applications.

as a starting point, we added another category of variables. For example, the first line of table 2 shows the results for a model specification where only socio-demographic variables are included. The results in line 5 are for a model with socio-demographic and qualification variables. Testing all possible combinations of two and three categories and finally using all information gives us 15 specifications for the four main groups. One shortcoming of such statistical tests becomes obvious from the results. If we take for example the results for women in West Germany we see that the best hit-rate (81.7%) is achieved by only including the regional dummy variables. With the full specification we achieve only a hit-rate of 75.7%. Following that rule would mean that we should estimate the propensity score solely based on the model with regional dummy variables. This makes no economic sense since obviously important characteristics, e.g. age and qualification variables, would be excluded. It has to be kept in mind that the main purpose of the propensity score estimation is not to predict selection into treatment as well as possible but to balance all covariates. Hence, we use the full specification for the estimation of the propensity scores. Table 3 contains the results of the propensity score estimation. Looking at this table clarifies that the influence of variables on the participation probability differs by regions and gender and highlights the appropriateness of the separate estimation of the propensity scores for men and women in West and East Germany.

INSERT TABLE 3 ABOUT HERE

We will use the estimated propensity scores in the following to implement the matching estimator.

4.2 Choosing the Matching Algorithm

After having specified the propensity score model, the next choice to be made concerns the matching algorithm to be used. Several algorithms have been suggested in the literature. Good overviews can be found in Heckman, Ichimura, Smith, and Todd (1998) or Smith and Todd (2005). Clearly, all approaches should yield asymptotically the same results, because with growing sample size all of them become closer to comparing only exact matches (Smith, 2000). However, in small samples the choice of the matching approach can be important (Heckman, Ichimura, and Todd, 1997). All matching estimators contrast the outcome of a treated individual with the outcome of comparison group members. However, the estimators differ not only in the definition of the neighbourhood for each treated individual and the handling of the common support problem, but also with respect to the weights given to these neighbours. Usually a trade-off between bias and variance arises. First, one has to decide on how many non-treated individuals to match to a single treated individual. Nearest-neighbour (NN) matching only uses the participant and its closest neighbour. Therefore it minimises the bias but might also involve an efficiency loss, since a large number of close neigbours is disregarded. Clearly, NN matching faces the risk of bad matches, if the closest neighbour is far away. This can be avoided by using caliper matching, i.e. imposing a tolerance on the maximum distance in the propensity scores allowed. Kernel-based matching on the other hand uses more (all) non-participants for each participant, thereby reducing the variance but possibly increasing the bias. Finally, using the same non-treated individual more than once (NN matching with replacement) can possibly improve the matching quality, but increases the variance.[18] Kernel matching is not feasible for our estimation, since the computing time is too high. However, to see if the inclusion of more comparison units for the construction of the counterfactual outcome has influence on the estimated effects, we also use 'oversampling' methods. This form of matching trades reduced variance, resulting from using more information to construct the counterfactual for each participant, with increased bias that results from poorer matches on average (Smith and Todd, 2005).[19]

Insert table 4 about here

This brief discussion makes clear that even with NN matching several alternatives emerge. It seems reasonable to try a number of approaches and test the sensitivity of the results with respect to the algorithm choice. If they give similar results, the choice may be unimportant. Else, if the results differ, further investigation may be needed in order to reveal more about the source of the disparity (Bryson,

[18] This is of particular interest with data where the distribution of the propensity score is very different in the treatment and the comparison group (see the discussion in Smith and Todd (2005)).

[19] When using oversampling, one has to decide how many matching partners m should be chosen for each individual i and which weight should be assigned to them. We will use uniform weights, that is all the m comparison individuals within set A_i receive the weight $\frac{1}{m}$, whereas all other individuals from the comparison group receive the weight zero.

Dorsett, and Purdon, 2002). We implement eleven matching algorithms, including NN matching without replacement (without caliper and with calipers of 0.01, 0.02 and 0.05) and NN matching with replacement with the same calipers. To see if the estimates differ when more neighbours are included, we additionally implement oversampling with 2, 5 and 10 nearest neighbours.[20]

Table 4 contains the results for the main groups for the last month of the observation period. Bold letters indicate significance at the 1%-level, italic letters refer to the 5%-level, standard errors are bootstrapped with 50 replications. The estimates illustrate two points: First of all, the results are not sensitive to the chosen matching algorithm. For men in West Germany the effects are insignificant and centered around zero. For men in East Germany the significant effects vary between -2.37% (5-NN-Matching) and -2.94% (NN with replacement). This means that the employment rate of men in East Germany, who started their JCS in February 2000, is in December 2002 on average between 2.37% and 2.94% lower when compared to matched non-participants. We will give an extensive interpretation of the results in section 5 and restrict the discussion here to sensitivity issues. The significant effects for females in East Germany vary between -1.06% (10 NN) and -1.93% (NN with replacement). The only group for whom a somewhat higher variation in the effects is detected are women in West Germany, where the lowest estimated effect is 4.51% (NN without replacement and without caliper) and the highest estimated effect is 6.1% (10 NN). The second point to note is that the standard errors are (as expected) in general lower for the oversampling algorithms, even though the differences here are not very pronounced. Hence, the choice of the matching algorithm seems not to be a critical issue in our case. The results show that the estimates are not sensitive to the algorithm choice and that the improvement which comes from oversampling methods in terms of reduced variance is limited only. Therefore, we decide to use NN matching for the further analysis. Since we have a very large sample of non-participants, the probability of finding good matches without using replacement is quite high. To avoid an unnecessary inflation of the variance, we match without replacement. Finally, to ensure

[20] Matching is implemented using the STATA code PSMATCH2 by Leuven and Sianesi (2003).

a good matching quality, we implement a caliper of 0.02.[21]

4.3 Common Support

Before assessing the matching quality, it is important to check the region of common support for participants and non-participants. The most straightforward way is a visual analysis of the density distribution of the propensity score in both groups.[22] It is a common finding that the distribution for non-participants is highly skewed to the left. Problems arise, when the distributions in both groups do not overlap. A good example are short-term unemployed men in West Germany, where quite a large amount of observations in the treatment group has a propensity score over 0.5 and nearly none of the comparison individuals can be found in this region. There are several ways of imposing the common support condition, e.g. by 'minima and maxima comparison' or 'trimming' (see Caliendo and Kopeinig (2005) for an overview). We impose the 'minima and maxima condition' and additionally implement NN matching with a caliper of 0.02. The idea of minima-maxima comparison is to delete all treated observations, whose propensity score is smaller than the minimum and higher than the maximum in the comparison group. Treated individuals who fall outside the common support region have to be disregarded and for these individuals the treatment effect cannot be estimated. Bryson, Dorsett, and Purdon (2002) note that if the proportion of lost individuals is small, this poses few problems. However, if the number is too large, there may be concerns whether the estimated effect on the remaining individuals can be viewed as representative.

Insert table 5 about here

Table 5 contains the number of treated individuals lost in each of the sub-groups. It can be seen that the number of lost individuals is fairly low for three of the main groups. For men in West Germany we lose 0.79% of the observations, for men (0.03%) and women (0.16%) in East Germany the proportion is even smaller. However, for women in West Germany we cannot find similar non-participants for

[21] This is mostly driven by the finding for women in West Germany, where imposing this caliper reduces the number of treated observations by approximately 4.6% of the sample. In turn, this means that if we do not impose this caliper, the distance in the propensity scores would be higher than 0.02 for 4.6%. For the other groups, imposing the caliper does not have much influence.

[22] The figures are available on request by the authors.

around 6.84% of the treated population and have to discard these individuals. The overlap between participating and non-participating women in West Germany is fairly limited in the sub-groups, too, and we lose up to 20.95% of the treated population (for women with placement restrictions). Hence, interpretation of the effects has to be made careful.

For the other sub-groups the share of lost individuals is acceptable. However, two sub-groups are problematic for both gender and regions. The first are short-term unemployed persons (less than 13 weeks unemployed). For this group we lose 21.15% of the participating men in West Germany, 19.20% of men and 25.04% of women in East Germany. This means that we are not able to find short-term unemployed individuals in the comparison group that have similar propensity scores as the treated individuals. The second sub-group are individuals with high degree, where the share of individuals lost is 6.85% (14.14%) for men (women) in East Germany and 14.29% (17.81%) for men (women) in West Germany. Overall, we note that the share of lost individuals is rather small in East Germany, higher for men in West Germany and highest for women in West Germany.

4.4 Matching Quality

Matching Quality for the Main Groups Since we do not condition on all covariates but on the propensity score, we have to check the ability of the matching procedure to balance the relevant covariates. One suitable indicator to assess the distance in the marginal distributions of the X-variables is the standardised bias (SB) suggested by Rosenbaum and Rubin (1985). For each covariate X it is defined as the difference of the sample means in the treated and (matched) comparison sub-samples as a percentage of the square root of the average of the sample variances in both groups. This is a common approach used in many evaluation studies, e.g. by Sianesi (2004). We estimate the absolute bias between the respective participating and non-participating groups before and after matching took place. To abbreviate the documentation, we calculated the means of the SB before and after matching for the four main groups (Table 6) as an unweighted average of all variables (mean standardised bias, MSB).[23] The overall bias before matching lies between 10.9% for women in East Germany and 15.36%

[23] The results for each variable are available on request by the authors.

for women in West Germany. A significant reduction can be achieved for all groups so that the bias after matching is 2.5% (3.1%) for men (women) in West Germany and 1.8% (1.6%) for men (women) in East Germany. Clearly, this is an enormous reduction and shows that the matching procedure is able to balance the characteristics in the treatment and the matched comparison group.

Insert table 6 about here

Additionally Sianesi (2004) suggests to re-estimate the propensity score on the matched sample (i.e. on participants and matched non-participants) and compare the pseudo-R^2's before and after matching. After matching there should be no systematic differences in the distribution of the covariates between both groups. Therefore, the pseudo-R^2 after matching should be fairly low. As the results from Table 6 show, this is true for our estimation. The results of the F-tests (with degrees of freedom in brackets) point in the same direction, indicating a joint significance of all regressors before, but not after matching.

Matching Quality for the Sub-Groups Now that we have shown that the matching procedure is able to balance the distribution of the covariates between treated and comparison individuals in the main groups, we have to test this for the sub-groups, too.

Insert table 7 about here

Table 7 contains the results for the eleven sub-groups. The first column refers to the MSB before matching, the second column to the MSB after matching, when matching is done with the estimated 'overall' propensity score as shown in table 3. This propensity score specification, which we label P_1, has been done separately for the four main groups. However, it is very clear that the matching procedure based on the overall scores is not able to balance the covariates between treated and matched non-treated individuals in the sub-groups. For example, the bias after matching for men in West Germany reaches a level of 13.23% for rehabilitation attendants. Even though this is a reduction compared to the MSB before matching, it is not acceptable. For women in West Germany in this group the bias after matching is 18.69%. In East Germany the bias after matching is not much lower, reaching levels of 13.11% for young men and 15.11% for young women. Even though there are some sub-groups for which the bias is acceptable, the overall matching quality in the sub-groups is not. Hence, alternative strategies are called for.

One way to do so is to re-define the propensity score estimation. Whereas the 'overall' propensity score estimation has only been done separately for men and women in West and East Germany, we also estimate 'group-specific' propensity scores. The basic idea behind that is to capture the varying influence of the variables for certain sub-groups more accurately. Since we have eleven sub-groups for both gender and regions, we are left with 44 propensity score estimations.[24] Based on these estimations, labelled P_2, we re-run the matching procedure and estimate the MSB once again. It can be seen that the MSB is now clearly lower not only compared to the situation before matching but also compared to the situation when matching on the 'overall' score. This result shows that using the 'overall' score specification has not been fine enough to balance the relevant characteristics between participants and non-participants in the sub-groups. Hence, we will use the 'group-specific' propensity scores for the further analysis in the sub-groups.

5 Empirical Results

An important decision which has to be made in every evaluation is when to measure the programme effects. The empirical analysis should ensure that participants and non-participants are compared in the same economic environment and the same lifecycle position. The literature is dominated by two approaches, either comparing individuals from begin or after the end of programmes. The latter approach is problematic for two reasons. First, since it implies comparison of participants and non-participants in the month(s) after programmes end, very different economic situations maybe compared if exits are spread over a longer time period. Second, this approach entails an endogeneity problem of programme exits (Gerfin and Lechner, 2002). A second approach which is predominant

[24] The results of these estimations are available on request by the authors.

in the recent literature (see e.g. Sianesi (2004) or Gerfin and Lechner (2002)) and which is also used here, measures the effects from begin of the programmes. By doing so, the policy-relevant question if the placement officer should place an unemployed individual in February 2000 in a JCS or not, can be answered. What should be kept in mind is the possible occurrence of locking-in effects for the group of participants. Van Ours (2004) notes that the net effect of a programme consists of two opposite effects: First, the employment probability of the participants is expected to rise due to positive aspects of the programme. Second, since participants who are involved in the programmes do not have the same time to look for new jobs as non-participants, a reduced search intensity during programmes is expected. Since it is not possible to disentangle both effects, locking-in effects should be seen as a constituent part of the overall programme effect (Sianesi, 2001). When interpreting the results the different impacts of the two underlying effects have to be considered. As to the fall in the search intensity, we should expect an initial negative locking-in effect from any kind of participation in a programme. To assess the possible magnitude of this initial effect it is helpful to look at the programme exit rates in each group.[25] Most of the participants leave the programmes after one year. In March 2001, around 80% (74%) of the male (female) participants in West Germany have left the programmes. The corresponding numbers are approximately 91% for men and 92% for women in East Germany. Since we observe the outcome of the individuals until almost three years after programmes start, successful programmes should overcompensate for this initial fall.

Results for the Main Groups The results from the begin (February 2000) to the end (December 2002) of our observation period for the main and the sub-groups are depicted in figures 1 to 4. Figure 1 contains the results for men in West Germany. The solid line in the graphs describes the monthly employment effect, i.e. the difference in the employment rates between participants and matched non-participants. The graphs for the main group are captioned 'total' in the figures. All of the graphs have one thing in common, namely a large drop in the effects for the first months after programme start. This can be interpreted as the expected locking-in effect, which is more pronounced for men (figure 1)

[25] Tables with cumulated exit rates for the main and sub-groups are available on request by the authors.

and women (figure 2) in West Germany than for men (figure 3) and women (figure 4) in East Germany. Five months after programmes have started (in July 2000), the effects for men in West Germany lie around -21.1%. That means that the average employment rate of participating men is about 21% lower in comparison to matched non-participants. Clearly, this strong reduction is expected as nearly all participants are still in the programmes, whereas the non-participants have the chance to search, apply for and find a new job. For the interpretation one has to bear in mind, that although JCS are some kind of employment, they are classified as failures when assessing the re-integration success into regular (unsubsidised) employment. For women in West Germany the result is very similar in that month and amounts to around -20.4%. The situation in East Germany is somewhat different. The effects are here -14.0% for men and -9.4% for women. Compared to the results for West Germany, this reflects the worse labour market situation with fewer employment opportunities. Being locked into the programme does not have as much influence, since the chances of non-participants to find a new job are lower anyway.

The development of the effects is quite different for both regions, too. Whereas in West Germany a relatively steep increase in the employment effects can be found, the development in East Germany is much smoother. For example, in July 2001 the employment effect has risen to -12.5% for men and -11.9% for women in West Germany. Hence, the negative effects are nearly halved. In East Germany, however, the effects lie around -10.9% for men and -7.5% for women. Looking at the last month of our observation period (December 2002), we do not find a significant programme effect for men in West Germany. That is, the employment chances of participants and matched non-participants do not differ. However, for women in West Germany we find a significant positive effect of 4.6%, which means that participating women have benefited from the programme in terms of employment chances. However, this positive result has to be treated with caution since women in West Germany have been the smallest group, we have lost a considerable share of participants due to the common support requirement and the estimates imply a confidence interval which is close to zero.

Insert figures 1-4 about here

For East Germany on the other hand, we find negative employment effects of -2.9% for men and -1.4% for women. This shows that the overall effect of JCS for the participating individuals is dissatisfying. Only for one of the groups, women in West Germany, we find a positive employment effect nearly three years after programmes have started, whereas for the other three main groups the effects are negative or insignificant. It seems that the pronounced initial negative (locking-in) effect cannot be overcome during our observation period. Judging by these numbers, JCS have to be rated as unsuccessful regarding their goal to re-integrate individuals into regular (unsubsidised) employment.

Results for the Selected Sub-Groups Even though JCS do not work for the participants as a whole, they may work for sub-groups. For instance, one could assume that they are especially effective for the explicit target groups of JCS, like long-term unemployed persons or persons without work experience. Figures 1 to 4 contain the results for our selected sub-groups. To abbreviate the discussion, we concentrate on two main points. First, we will examine the occurrence of locking-in effects and second, we will discuss the results at the end of the observation period (December 2002). Considering locking-in effects is explicitly of interest, since it can be expected that these effects differ for the sub-groups. Good examples are provided by the groups defined by age and unemployment duration. Older unemployed persons have in general fewer labour market opportunities than middle-aged or younger persons. Due to the worse 'outside options' of the non-participants, we expect to find weaker locking-in effects for older participants and stronger effects for the other groups. The figures support these expectations empirically, independently of gender of region. With regard to the previous unemployment duration, it can be assumed that re-integration into the labour market is generally easier for persons with only a short duration of unemployment ('negative duration dependence'). Therefore, short-term unemployed non-participants are expected to have a higher probability of receiving a job offer and hence the locking-in effects should be larger which is in fact the case. For the other sub-groups the graphs present a similar picture. We find the initial fall of the employment effects in the first months after programmes have started and rising tendencies after the majority of participants has left the programmes.

The second point we want to discuss are the effects for these sub-groups at the end of our observation period (December 2002). For most of the groups we do not find significant programme effects at this point in time, i.e. the employment rates of participants and matched non-participants do not differ nearly three years after programmes have started. That implies that programmes have neither improved nor worsened the employment chances of participating individuals. However, for some of the groups we find significant differences in the employment rates. Long-term unemployed (more than 52 weeks) men (5.0%) and women (11.3%) in West Germany benefit from participation. These results indicate that JCS could improve the employment chances of this target group. Additionally, high qualified men in West Germany benefit from participation (12.5%), whereas for low qualified persons and individuals without work experience no significant effects can be established. This is intuitively not understandable, since programmes should be designed for persons who are most in need of assistance. Another group who benefits from participation are older women in West Germany, whose employment rate is 12.7% higher than for matched non-participants. This is an encouraging result, because older unemployed persons in particular have only poor opportunities to return to the first labour market. Although for most groups we do not find any enhancement of the employment chances after participation, the results indicate a tendency that programmes are actually only useful for the most-disadvantaged in terms of unemployment duration and age.

Considering the results for the sub-groups in East Germany reveals a somewhat different picture. Focussing on the male groups, we only find a significant negative effect (-10.1%) for participants with a short unemployment duration before programme. What has to be kept in mind, is the relatively high share of individuals which we have lost due to the common support restriction. Hence, we are reluctant to overemphasise the relevance of this effect. For all other groups no significant differences in the employment rates can be established. For women in East Germany the results are disappointing as well. Middle-aged (-2.2%) as well as short-term unemployed women (-7.4%) suffer from participation. Another group with clearly negative programme effects in December 2002 are high qualified women

(-9.8%). However, there is also one group (long-term unemployed women) for whom we find a small (2.5%) positive programme effect. For the other groups no significant differences can be established. Thus, the above stated hypothesis that programmes are actually only likely to work for the legally defined target groups, can only be supported for long-term unemployed women.

Sensitivity of the Results to Unobserved Heterogeneity The estimation of treatment effects with matching estimators is based on the CIA. Hence, if both groups differ on unobserved variables which simultaneously affect assignment into treatment and outcomes, there may be a 'hidden bias' to which matching estimators are not robust. Since it is impossible to estimate the magnitude of selection bias with non-experimental data, we address this problem with the bounding approach proposed by Rosenbaum (2002). The basic question to be answered is, if inference about programme effects may be altered by unobserved factors. In other words, we want to determine how strongly an unmeasured variable must influence the selection process in order to undermine the implications of matching analysis. Recent applications of this approach can be found in Aakvik (2001), DiPrete and Gangl (2004) and Hujer, Caliendo, and Thomsen (2004).

We outline this approach briefly, an extensive discussion can be found in Rosenbaum (2002). The participation probability for individual i with observed characteristics x_i can be written as $P(x_i) = P(D_i = 1 \mid x_i) = F(\beta x_i + \gamma u_i)$, where u_i is the unobserved variable and γ is the effect of u_i on the participation decision. Clearly, if there is no hidden bias, γ will be zero and the participation probability will solely be determined by x_i. However, if hidden bias exists, two individuals with the same observed covariates x have differing chances of receiving treatment. Following Aakvik (2001), we assume for the sake of simplicity that the unobserved covariate is a dummy variable with $u_i \in \{0, 1\}$. To give an example, u_i may represent individual motivation where a person is either motivated ($u = 1$) or not ($u = 0$). Rosenbaum (2002) derives the following bounds on the log-odds ratio that either of the two matched individuals will receive treatment:

$$\frac{1}{e^{\gamma}} \leq \frac{P(x_i)(1 - P(x_j))}{P(x_j)(1 - P(x_i))} \leq e^{\gamma}. \tag{3}$$

Matched individuals have the same probability of participating only if $e^\gamma = 1$. Else, if for example $e^\gamma = 2$, individuals who appear to be similar (in terms of x) may differ in their odds of receiving the treatment by as much as a factor of 2. In this sense, e^γ is a measure of the degree of departure from an estimation that is free of hidden bias (Rosenbaum, 2002). The basic idea now is to increase the influence of e^γ and see if inference from the test statistic is changed. Aakvik (2001) suggests to use the Mantel and Haenszel (1959) test statistic, which compares the successful number of persons in the treatment group against the same expected number, given the treatment effect is zero.[26]

Table 8 contains the results of the sensitivity analysis for two selected months (July and December 2002) of the examined sub-groups. First of all, the table contains the effects and the results of the Mantel and Haensel test-statistic for the situation free of hidden bias ($e^\gamma = 1$). A χ^2-value below 3.84 indicates that the treatment effect is insignificant (5% level). Clearly, a sensitivity analysis for insignificant effects is not meaningful and hence will be omitted. For the significant effects, we gradually increase the level of e^γ until the inference about the treatment effect is changed. In other words, we are assessing the strength unmeasured influences would require in order to change inference about the treatment effect. The interpretation is straightforward: Taking the effect for men in West Germany in July 2002 as an example, we see that the effect is -3.06% and significant. The critical value of e^γ is between 1.50 and 1.55. A critical value of 1.50 suggests that individuals with the same X-vector differ in their odds of participation by a factor of 1.50, or 50%. It is important to note that these are worst-case scenarios. Hence, a critical value of $e^\gamma = 1.50$ does not mean that unobserved heterogeneity exists and that there is no effect of treatment on the outcome variable. This result only states that the confidence interval for the effect would include zero if an unobserved variable caused the odds ratio of treatment assignment to differ between treatment and comparison groups by 1.50. Additionally, this variable's effect on the outcome would have to be so strong that it almost

[26] The Q_{MH} test-statistic follows the chi-square distribution with one degree of freedom and is given by:

$$Q_{MH} = \frac{U^2}{Var(U)} = \frac{[\sum_{s=1}^{S}(y_{1s} - \frac{n_{1s}y_s}{n_s})]^2}{\sum_{s=1}^{S}\frac{n_{1s}n_{0s}y_s(n_s-y_s)}{n_s^2(n_s-1)}}.$$

perfectly determines the outcome in each pair of matched cases in the data. However, even if there is unobserved heterogeneity to a degree of $e^{\gamma} = 1.50$ in the group of West German men, inference about the treatment effect would not be changed.

Insert table 8 about here

The results are ambivalent and differ between West and East Germany. In West Germany most of the effects for men and women in the sub-groups are insignificant. But for those groups where the effects are significant, even a large influence of unobserved heterogeneity does not have much influence on the inference about treatment effects. The lowest critical value of e^{γ} can be found for men without professional training in July 2002 (1.40-1.45) and the largest critical values of 1.75-1.80 can be found for high-qualified men (in July and December 2002) as well as for long-term unemployed women in July 2002 and high qualified women in December 2002. Therefore, we can conclude for West Germany that even large amounts of unobserved heterogeneity would not alter the inference about the estimated effects.

In contrast to that the results in East Germany are not so clear-cut. We find that for some of the sub-groups, like older or short-term unemployed men as well as for high qualified women and for women with placement restrictions, inference would change in July 2002 even with small amounts of hidden bias. The critical value of e^{γ} is somewhere below 1.05, which implies that even small magnitudes of 'hidden bias' would alter the inference. Consequently, interpretation for these sub-populations hinges on this restriction. For the results of the main groups, i.e. men and women, the critical value of e^{γ} in December 2002 is somewhere between 1.25 and 1.30. So these effects can be viewed as relatively robust to unobserved heterogeneity.

6 Conclusion

The aim of this paper is the evaluation of the re-integration effects of JCS into regular (unsubsidised) employment for the participating individuals in Germany. Our analysis is based on a dataset from administrative sources of the FEA containing information on all participants (11,151) who started a JCS in February 2000 and a comparison group of 219,622 unemployed persons. Special attention is given to the possible occurrence of individual, i.e. group-specific, and regional effect heterogeneity. That is we estimate the effects separately for men and women in West and East Germany ('main groups') as well as for eleven 'sub-groups'. Given the very informative dataset, we can apply a matching estimator to solve the problem of selection bias. Since the large number of relevant covariates makes exact matching unfeasible, we use propensity score matching for the analysis and discuss the required implementation steps in detail. We estimate the effects from begin of the programmes in February 2000 until December 2002. Since JCS are usually supported for twelve months, we find large locking-in effects for all of the groups in the first months after programme begin. The locking-in effects are more pronounced in West Germany and less substantial in East Germany, which may be caused by the better employment opportunities for non-participants in the West.

Regarding the effects for the main groups at the end of the observation period, we find a significant positive effect only for women in West Germany (4.6%), whereas the effect for men in West Germany is insignificant. For men (-2.9%) and women (-1.4%) in East Germany the effects are significantly negative. Hence, except for women in West Germany, it seems that the initial negative locking-in effect cannot be overcome during the observation period. For most of the sub-groups we do not find significant effects at all. However, one exception has to be noted. Long-term unemployed men (5.0%) and women (12.7%) in West Germany as well as long-term unemployed women (2.5%) in East Germany benefit from participation. The positive findings for the long-term unemployed persons indicate that JCS do work for this problem group of the labour market. However, this result cannot be extended to other problem groups, like individuals with placement restrictions, individuals without work experience or low qualified persons. Even though we would have expected positive effects for these problem groups of the labour market, we did not find any. To some extent the effects reflect the different purpose of JCS in both parts. Whereas they are used as a relief of the labour market in East Germany they are more tightly addressed to problem groups in the West leading to better effects.

The overall picture is rather disappointing since most of the effects are insignificant or negative. Participation in programmes does not help individuals to re-integrate into regular (unsubsidised) employment. The results are concordant with recent evaluation studies of JCS for other countries (see e.g. Sianesi (2004) for Sweden or Martin and Grubb (2001) for an overview of OECD countries), finding large locking-in effects and overall negative effects. However, our results also emphasise the importance of effect heterogeneity when estimating treatment effects. We find positive effects for long-term unemployed, who are usually one of the most problematic groups on the labour market with only the slightest re-employment chances. Hence, the positive result for them is promising and shows that JCS work for this target group and might be an alternative for hard-to-place individuals. Clearly, one policy implication is to address programmes to this problem group more tightly, which is at the moment, especially in East Germany, not the case. Limiting access to these programmes and tailoring them more for the ones who need them most might be a way to improve their overall efficiency and offering a 'last chance' for hard-to-place individuals.

References

Aakvik, A. (2001): "Bounding a Matching Estimator: The Case of a Norwegian Training Program," *Oxford Bulletin of Economics and Statistics*, 63(1), 115–143.

Angrist, J. D., and A. B. Krueger (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1277–1366. Elsevier Science B.V.

Blien, U., F. Hirschenauer, M. Arendt, H. J. Braun, D.-M. Gunst, S. Kilcioglu, H. Kleinschmidt, M. Musati, H. Ross, D. Vollkommer, and J. Wein (2004): "Typisierung von Bezirken der Agenturen der Arbeit," *Zeitschrift für Arbeitsmarktforschung*, 37(2), 146–175.

Blundell, R., and M. Costa Dias (2002): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal*, 1, 91–115.

Breiman, L., J. Friedman, R. Olsen, and C. Stone (1984): *Classification and Regression Trees*. Wadsworth International Group, Belmont.

Bryson, A., R. Dorsett, and S. Purdon (2002): "The Use of Propensity Score Matching in the Evaluation of Labour Market Policies," Working Paper No. 4, Department for Work and Pensions.

Bundesministerium für Wirtschaft und Arbeit (2003): "Moderne Dienstleistungen am Arbeitsmarkt," http://www.bmwa.bund.de/Navigation/Arbeit/Arbeitsmarktpolitik/moderne-dienstleistungen-am-arbeitsmarkt,did=10144.html.

Caliendo, M., R. Hujer, and S. Thomsen (2005): "Identifying Effect Heterogeneity to Improve the Efficiency of Job Creation Schemes in Germany," Discussion Paper No. 05-21, ZEW, Mannheim.

Caliendo, M., and S. Kopeinig (2005): "Some Practical Guidance for the Implementation of Propensity Score Matching," Discussion Paper No. 1588, IZA, Bonn.

Dawid, A. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, 1–31.

DiPrete, T., and M. Gangl (2004): "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments," Working Paper, WZB.

Fisher, R. (1935): *Design of Experiments*. Hafner, New York.

Gerfin, M., and M. Lechner (2002): "A Microeconometroc Evaluation of the Active Labour Market Policy in Switzerland," *The Economic Journal*, 112, 854–893.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heckman, J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics Vol.III*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. Elsevier, Amsterdam.

Heckman, J., and J. Smith (1999): "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies," Working Paper No. 6983, National Bureau of Economic Research.

Hujer, R., M. Caliendo, and S. Thomsen (2004): “New Evidence on the Effects of Job Creation Schemes in Germany - A Matching Approach with Threefold Heterogeneity,” *Research in Economics*, 58(4), 257–302.

Imbens, G. (2004): “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review,” *The Review of Economics and Statistics*, 86(1), 4–29.

Leuven, E., and B. Sianesi (2003): “PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing,” Software, http://ideas.repec.org/c/boc/bocode/s432001.html.

Mantel, N., and W. Haenszel (1959): “Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease,” *Journal of the National Cancer Institute*, 22, 719–748.

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

Neyman, J. (1935): “Statistical Problems in Agricultural Experiments,” *The Journal of the Royal Statistical Society*, 2, 107–180.

Puhani, P. (1998): “Advantage through Training? A Microeconometric Evaluation of the Employment Effects of Active Labour Market Programmes in Poland,” Discussion Paper No. 98-25, ZEW.

Quandt, R. (1972): “Methods for Estimating Switching Regressions,” *Journal of the American Statistical Association*, 67(338), 306–310.

——— (1988): *The Economics of Disequilibrium*. Basil Blackwell, Oxford.

Rosenbaum, P., and D. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–50.

——— (1985): “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,” *The American Statistican*, 39, 33–38.

Rosenbaum, P. R. (2002): *Observational Studies*. Springer, New York.

Roy, A. (1951): “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135–145.

Rubin, D. (1974): “Estimating Causal Effects to Treatments in Randomised and Nonrandomised Studies,” *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B., and N. Thomas (1996): “Matching Using Estimated Propensity Scores: Relating Theory to Practice,” *Biometrics*, 52, 249–264.

Sianesi, B. (2001): “Differential effects of Swedish active labour market programmes for unemployed adults during the 1990s,” Working Paper No. 01/25, The Institute for Fiscal Studies.

——— (2004): “An Evaluation of the Active Labour Market Programmes in Sweden,” *The Review of Economics and Statistics*, 86(1), 133–155.

Smith, H. (1997): “Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies,” *Sociological Methodology*, 27, 325–353.

Smith, J. (2000): “A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies,” *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 136(3), 1–22.

Smith, J., and P. Todd (2005): “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?,” *Journal of Econometrics*, 125(1-2), 305–353.

van Ours, J. (2004): “The Locking-in Effect of Subsidized Jobs,” *Journal of Comparative Economics*, 32(1), 37–52.

Tables and Figures

Tab. 1: Number of Observations in Main and Sub-Groups

	West Germany				East Germany[1]			
	Men		Women		Men		Women	
Groups	Part.	Non-Part.	Part.	Non-Part.	Part.	Non-Part.	Part.	Non-Part.
Total (Main group)	2,140	44,095	1,052	34,227	2,924	64,788	5,035	76,512
Sub-groups								
Age (in years)								
<25 years	458	4,102	182	2,443	240	8,743	148	4,864
25-50 years	1,337	23,560	709	19,732	1571	35,927	3,342	44,329
>50 years	345	16,433	161	12,052	1,113	20,118	1,545	27,319
Unemployment duration (in weeks)								
<13 weeks	558	12,198	237	7561	578	22,003	575	12,447
13-52 weeks	744	13,909	403	12,235	1,248	22,864	1,970	26,657
>52 weeks	838	17,988	412	14,431	1,098	19,921	2,490	37,408
Without professional experience	273	3,281	159	2,548	293	7,023	498	7945
Without professional training	1,340	21,659	476	17,093	837	14,966	1,121	19,776
With high degree	112	1,486	146	1,165	146	2,682	191	1,619
Rehabilitation attendant	111	2,763	44	1,063	218	4,849	156	3,520
With placement restrictions	354	9,516	148	5,993	394	10,470	376	9,121

[1] Observations from the labour office districts of Berlin are excluded.

Tab. 2: Hit-Rates and Pseudo-R^2 for Different Propensity Score Specifications [1]

Specification (Sets of Variables included)				West Germany				East Germany			
				Men		Women		Men		Women	
Socio-Demogr. [2]	Qualification [3]	Career [4]	Region [5]	Hit-Rate	R^2	Hit-Rate	R^2	Hit-Rate	R^2	Hit-Rate	R^2
x				55.20	0.036	64.51	0.050	44.46	0.014	54.47	0.019
	x			61.03	0.033	76.19	0.036	59.10	0.014	67.29	0.013
		x		73.21	0.106	79.45	0.130	76.64	0.106	72.70	0.097
			x	54.00	0.000	81.67	0.001	63.69	0.000	73.95	0.000
x	x			62.60	0.062	67.29	0.076	58.12	0.030	55.18	0.030
x		x		68.81	0.122	73.03	0.153	74.56	0.116	72.20	0.105
x			x	55.21	0.036	65.31	0.051	45.64	0.014	54.11	0.019
	x	x		69.74	0.123	78.09	0.153	75.36	0.110	72.61	0.106
	x		x	61.40	0.033	74.63	0.037	55.98	0.014	67.60	0.013
		x	x	72.62	0.106	77.01	0.133	76.26	0.106	72.81	0.098
x	x	x		70.65	0.138	75.53	0.174	74.28	0.122	72.18	0.113
x	x		x	62.61	0.062	67.26	0.077	57.94	0.030	55.28	0.030
x		x	x	68.97	0.123	73.23	0.157	74.51	0.117	72.12	0.106
	x	x	x	69.99	0.124	77.51	0.156	75.12	0.111	72.69	0.107
x	x	x	x	70.60	0.139	75.70	0.177	74.20	0.122	72.24	0.114

[1] *Hit-rates* are computed in the following way: If the estimated propensity score is larger than the sample proportion of persons taking treatment, i.e. $\hat{P}(X) > \overline{P}$, observations are classified as '1'. If $\hat{P}(X) \leq \overline{P}$ observations are classified as '0'.
[2] *Socio-demographic variables* include age, age^2, marital status, number of children, nationality and health restrictions.
[3] *Qualification variables* include professional training, occupational group, professional rank and work experience.
[4] *Career variables* include duration of last employment and unemployment, number of placement propositions, last contact to job center, rehabilitation attendants, placement restrictions and previous labour market programmes.
[5] *Regional variables* consist of seven FEA clusters as defined in Blien *et al.*(2004).

Tab. 3: Estimation Results of the Logit-Models for the Propensity Score

	West Germany				East Germany			
	Men		Women		Men		Women	
Variable	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Constant	**-1.1739**	0.2731	**-3.1254**	0.4533	**-5.7880**	0.3659	**-8.0021**	0.3944
Socio-Demographic Variables								
Age	**-0.0599**	0.0145	-0.0067	0.0235	**0.0901**	0.0141	**0.1702**	0.0136
Age^2	*0.0004*	0.0002	-0.0003	0.0003	**-0.0008**	0.0002	**-0.0019**	0.0002
Married	**-0.1676**	0.0612	**-0.4483**	0.0761	**0.2683**	0.0506	**0.1145**	0.0344
Number of children	*0.0653*	0.0281	-0.0183	0.0439	-0.0335	0.0266	-0.0238	0.0184
German	**0.4402**	0.0683	*0.2825*	0.1211	**0.6284**	0.1966	**0.7082**	0.2432
Health restrictions								
No health restrictions	Ref.		Ref.		Ref.		Ref.	
Acc. DoR[1], 80% and over	**0.9160**	0.1826	**1.3404**	0.2578	*0.5491*	0.2758	**1.1375**	0.2442
Acc. DoR, 50% to under 80%	**0.8052**	0.1267	**0.6433**	0.1978	**0.4991**	0.1270	**0.6032**	0.1242
Acc. DoR, 30% to under 50%	**1.1190**	0.3658	**1.9871**	0.4246	**0.5691**	0.1925	**0.7999**	0.1954
Acc. DoR, 30% to under 50%, no equalis.[2]	0.2757	0.1570	0.0651	0.2685	-0.0708	0.1721	-0.0725	0.1826
Other health restrictions	-0.0472	0.0892	-0.0751	0.1390	**-0.1918**	0.0716	*-0.1422*	0.0608
Qualification Variables								
Professional training								
Without compl. prof. training, no CSE[3]	Ref.		Ref.		Ref.		Ref.	
Without compl. prof. training, with CSE	**-0.3364**	0.0622	0.2294	0.1334	0.1015	0.0823	**0.3428**	0.0865
Industrial training	**-0.6738**	0.0692	-0.0808	0.1399	*-0.1777*	0.0748	**0.3315**	0.0820
Full-time vocational school	**-0.7639**	0.2685	-0.0734	0.2432	-0.3223	0.2594	**0.8588**	0.1384
Technical school	-0.0987	0.1756	**0.7183**	0.1927	0.2227	0.1231	**1.0166**	0.0977
Polytechnic	0.3534	0.2009	**1.4983**	0.2144	-0.0135	0.2058	**1.0388**	0.1794
College, university	0.2399	0.1577	**1.0221**	0.1869	0.0810	0.1354	**0.9004**	0.1272
Occupational group								
Plant cultivation, breeding, fishery	*0.2222*	0.0927	0.2628	0.2501	0.0092	0.0828	**0.2370**	0.0670
Mining, mineral extraction	-0.5605	0.4657			-0.7494	0.5154		
Manufacturing	Ref.		Ref.		Ref.		Ref.	
Technical professions	**-0.5810**	0.1544	-0.1609	0.2605	-0.1954	0.0999	**0.2149**	0.0819
Service professions	**-0.3077**	0.0544	**0.3167**	0.0995	**-0.1739**	0.0478	0.0127	0.0406
Other professions	0.1023	0.1533	0.3933	0.2628	**-1.1891**	0.2170	**-1.2092**	0.2860
Professional rank								
Unskilled worker	Ref.		Ref.		Ref.		Ref.	
Skilled worker	**-0.5499**	0.0982	-0.1637	0.1944	**-0.1811**	0.0597	0.0657	0.0525
White-collar worker, simple occupations	0.0163	0.1152	0.1490	0.1256	0.1809	0.1067	**0.2197**	0.0605
White-collar worker, advanced occupations	0.0877	0.1536	**0.5131**	0.1624	-0.2838	0.1662	-0.0404	0.1215
Other	-0.0112	0.0563	0.1512	0.1054	0.0345	0.0528	*0.1004*	0.0437
Qualification (with work experience)	**-0.3397**	0.0745	**-0.3139**	0.1017	**-0.2279**	0.0695	*-0.1175*	0.0527
Career Variables								
Duration of last employment (months)	**-0.0046**	0.0005	**-0.0033**	0.0007	**-0.0038**	0.0004	**-0.0028**	0.0003
Duration of unemployment (weeks)								
Up to 13 weeks	Ref.		Ref.		Ref.		Ref.	
Between 13 and 52 weeks	**0.2055**	0.0616	0.0698	0.0889	**0.4673**	0.0561	**0.2509**	0.0511
More than 52 weeks	**0.3087**	0.0678	0.0888	0.0974	**0.4498**	0.0599	**0.1694**	0.0509
Number of placement propositions	**0.0494**	0.0028	**0.0530**	0.0042	**0.0610**	0.0030	**0.0919**	0.0031
Last contact to job center (weeks)	-0.0013	0.0125	**0.0520**	0.0177	**-0.1204**	0.0114	**-0.0644**	0.0085
Rehabilitation attendant	-0.1533	0.1185	0.0696	0.2039	**0.2958**	0.0939	0.1535	0.1024
Placement restrictions	**-0.3396**	0.0989	-0.2654	0.1546	**-0.3164**	0.0870	**-0.3000**	0.0825
Programme before unemployment								
No further education or programme	Ref.		Ref.		Ref.		Ref.	
Further education compl., cont. education	**0.2292**	0.0801	**0.5301**	0.1043	**0.4830**	0.0628	**0.5263**	0.0422
Further education compl., voc. adjustment	**0.6479**	0.2286	0.4613	0.4466	**0.6545**	0.0893	**0.5634**	0.0746
Job-preparative measure	-0.4764	1.0285	**2.6387**	0.5245	**1.1431**	0.4289	0.3364	0.5250
Job creation scheme	**2.1463**	0.0777	**3.0671**	0.1141	**1.7272**	0.0546	**1.5382**	0.0418
Rehabilitation measure	-0.0929	0.2706	**0.9368**	0.3406	0.4232	0.2273	0.3780	0.2720
Regional Context Variables[4]								
Cluster Ia					-0.1040	0.1291	0.1421	0.1238
Cluster Ib					*-0.3077*	0.1248	-0.0242	0.1210
Cluster Ic					*-0.2838*	0.1361	-0.1841	0.1292
Cluster II	**-0.2225**	0.0730	**-0.5666**	0.0960	Ref.		Ref.	
Cluster III	*-0.1841*	0.0722	**-0.4601**	0.0917				
Cluster IV	-0.0080	0.1002	**-0.4530**	0.1423				
Cluster V	Ref.		Ref.					

Bold letters indicate significance at the 1% level. *Italic* letters refer to the 5% level.

[1] DoR = degree of restriction.

[2] People with accepted degree of restriction, but no equalisation to other persons with the same DoR.

[3] CSE = Certificate for secondary education

[4] Cluster according to the FEA classification as described in Blien *et al.*(2004).

Tab. 4: The Effects in the Main Groups for Different Matching Algorithms[1,2]

West Germany	**Men**			**Women**		
Matching Algorithm	Effect	S.E.	Obs.[3]	Effect	S.E.	Obs.[3]
NN without replacement	-0.0005	0.0108	2,132	**0.0554**	0.0200	1,028
caliper 0.01	-0.0028	0.0137	2,119	*0.0451*	0.0213	975
caliper 0.02	-0.0019	0.0158	2,123	0.0459	0.0258	980
caliper 0.05	-0.0009	0.0128	2,131	*0.0479*	0.0223	1,002
NN with replacement	0.0061	0.0110	2,140	*0.0504*	0.0231	1,052
caliper 0.01	0.0042	0.0139	2,132	*0.0504*	0.0233	1,051
caliper 0.02	0.0056	0.0150	2,139	*0.0504*	0.0211	1,052
caliper 0.05	0.0061	0.0133	2,140	*0.0504*	0.0207	1,052
Oversampling						
2 NN	0.0023	0.0140	2,140	*0.0466*	0.0221	1,052
5 NN	0.0011	0.0106	2,140	**0.0529**	0.0161	1,052
10 NN	-0.0003	0.0100	2,140	**0.0610**	0.0180	1,052
East Germany	**Men**			**Women**		
Matching Algorithm	Effect	S.E.	Obs.[3]	Effect	S.E.	Obs.[3]
NN without replacement	**-0.0291**	0.0080	2,924	*-0.0135*	0.0075	5,032
caliper 0.01	**-0.0289**	0.0085	2,908	*-0.0137*	0.0070	5,026
caliper 0.02	**-0.0287**	0.0088	2,923	*-0.0135*	0.0064	5,027
caliper 0.05	**-0.0291**	0.0101	2,924	-0.0135	0.0076	5,027
NN with replacement	**-0.0294**	0.0112	2,924	**-0.0193**	0.0063	5,035
caliper 0.01	**-0.0294**	0.0086	2,924	**-0.0191**	0.0069	5,031
caliper 0.02	**-0.0294**	0.0092	2,924	**-0.0193**	0.0075	5,032
caliper 0.05	**-0.0294**	0.0105	2,924	*-0.0193*	0.0081	5,034
Oversampling						
2 NN	**-0.0250**	0.0090	2,924	-0.0128	0.0073	5,035
5 NN	**-0.0237**	0.0065	2,924	-0.0101	0.0055	5,035
10 NN	**-0.0249**	0.0076	2,924	**-0.0106**	0.0038	5,035

Bold letters indicate significance at the 1%-level, *italic* letters refer to the 5%-level. Standard errors are bootstrapped with 50 replications.

[1] Nearest Neighbour (NN) matching without replacement uses each non-participant only once, whereas with NN matching with replacement each non-participant can be used repeatedly. Caliper defines the maximal allowed difference in the propensity score of participants and matched non-participants.

[2] Matching is implemented with the Stata module PSMATCH2 by Leuven and Sianesi (2003).

[3] Obs. is the number of participants after matching.

Tab. 5: Number of Treated Individuals Lost Due to Common Support Requirement[1,2]

West Germany	**Men**			**Women**		
	Before	After	Lost	Before	After	Lost
	Matching		in %	Matching		in %
Total	2140	2123	0.79	1052	980	6.84
Age (in years)						
<25	458	434	5.24	182	162	10.99
25-50	1337	1328	0.67	709	663	6.49
>50	345	344	0.29	161	150	6.83
Duration of unemployment						
< 13 weeks	558	440	21.15	237	189	20.25
13-52 weeks	744	720	3.23	403	365	9.43
> 52 weeks	838	835	0.36	412	400	2.91
Without professional experience	273	247	9.52	159	128	19.50
Without professional training	1340	1296	3.28	476	447	6.09
With high degree	112	96	14.29	146	120	17.81
Rehabilitation attendant	111	100	9.91	44	35	20.45
With placement restrictions	354	326	7.91	148	117	20.95
East Germany	**Men**			**Women**		
	Before	After	Lost	Before	After	Lost
	Matching		in %	Matching		in %
Total	2924	2923	0.03	5035	5027	0.16
Age (in years)						
<25	240	229	4.58	148	144	2.70
25-50	1571	1570	0.06	3342	3335	0.21
>50	1113	1074	3.50	1545	1481	4.14
Duration of unemployment						
< 13 weeks	578	467	19.20	575	431	25.04
13-52 weeks	1248	1230	1.44	1970	1963	0.36
> 52 weeks	1098	1098	0.00	2490	2490	0.00
Without professional experience	293	289	1.37	498	489	1.81
Without professional training	837	835	0.24	1121	1116	0.45
With high degree	146	136	6.85	191	164	14.14
Rehabilitation attendant	218	215	1.38	156	148	5.13
With placement restrictions	394	371	5.84	376	362	3.72

[1] We used the minima-maxima restriction as common support condition.
[2] Results refer to a NN matching without replacement and a caliper of 0.02.

Tab. 6: Some Quality Indicators

	West Germany		**East Germany**	
	Men	Women	Men	Women
Before Matching				
Pseudo R^2	0.1389	0.1775	0.1225	0.1144
F-Test[1]	2,406.8	1,679.4	2,951.3	4,323.3
Mean of standardised bias[2]	14.62	16.08	12.01	10.83
After Matching				
Pseudo-R^2	0.006	0.009	0.004	0.003
F-Test[1]	38.0	23.4	35.3	39.2
Mean of standardised bias[2]	2.51	3.20	1.78	1.60

[1] Degrees of freedom for the F-Test: 41 for men and 40 for women.
[2] Mean standardised bias has been calculated as an unweighted average of all covariates.

Tab. 7: Mean Standardised Bias in the Sub-Groups[1,2,3]

	West Germany						East Germany					
	Men			Women			Men			Women		
	Before Matching	After Matching with		Before Matching	After Matching with		Before Matching	After Matching with		Before Matching	After Matching with	
Propensity Score Specification		P_1	P_2		P_1	P_2		P_1	P_2		P_1	P_2
Sub-Groups												
Age (in years)												
<25	10.48	11.53	3.08	12.50	14.37	6.82	14.74	13.11	4.94	13.73	15.11	8.90
25-50	11.30	5.82	2.66	15.56	5.79	2.98	11.91	3.98	2.47	9.84	2.48	1.36
>50	17.82	12.48	5.83	20.48	12.70	6.62	16.79	9.48	2.55	14.98	6.56	1.55
Duration of unemployment												
< 13 weeks	15.71	9.79	4.96	16.18	12.19	4.58	19.05	10.47	4.74	13.78	10.80	4.37
13-52 weeks	12.79	5.89	4.04	16.10	7.51	4.32	12.43	3.45	2.42	11.74	4.03	1.53
> 52 weeks	17.77	6.03	3.06	19.13	7.65	4.18	13.55	7.51	2.00	11.61	3.48	1.69
Without professional experience	14.02	9.50	5.69	15.93	11.49	6.36	12.10	10.43	4.18	12.17	5.51	3.35
Without professional training	14.31	3.99	3.29	16.79	4.68	4.25	11.17	5.68	2.48	11.04	4.42	2.72
With high degree	17.18	10.64	7.52	14.50	9.64	5.77	18.14	12.99	6.10	15.04	12.48	5.64
Rehabilitation attendant	18.13	13.23	8.45	23.96	18.69	16.31	12.88	10.05	4.38	15.87	10.71	5.87
With placement restrictions	19.29	8.92	4.61	26.99	11.37	4.99	15.35	8.22	3.91	18.37	6.84	3.11

Standardised before matching calculated as: $100 \cdot (\overline{X}_1 - \overline{X}_0)/\{\sqrt{(V_1(X) + V_0(X))/2}\}$.

Standardised after matching calculated as: $100 \cdot (\overline{X}_{1M} - \overline{X}_{0M})/\{\sqrt{(V_{1M}(X) + V_{0M}(X))/2}\}$.

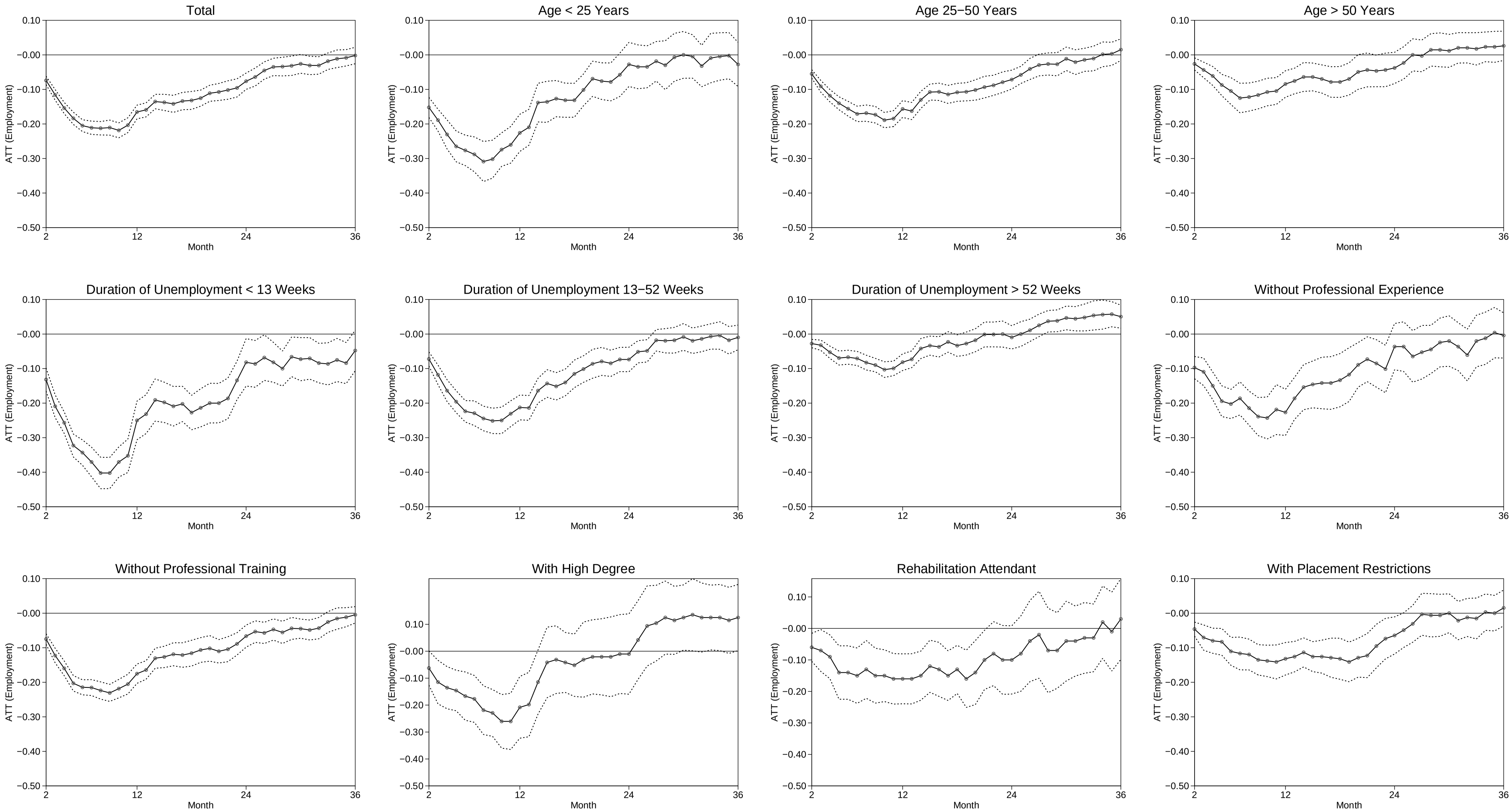
[1] Mean standardised bias has been calculated as an unweighted average of all covariates.
[2] P_1 refers to the 'overall' propensity score estimation.
[3] P_2 refers to the 'group-specific' propensity score estimation.

Tab. 8: Sensitivity Analysis for Unobserved Heterogeneity

	July 2002			December 2002			July 2002			December 2002		
West Germany	Men						Women					
Group	Effect	Q-MH for $e^\gamma = 1$	Critical value for e^γ	Effect	Q-MH for $e^\gamma = 1$	Critical value for e^γ	Effect	Q-MH for $e^\gamma = 1$	Critical value for e^γ	Effect	Q-MH for $e^\gamma = 1$	Critical value for e^γ
Total	*-0.0306*	5.35	1.50-1.55	-0.0019	n.s	n.s.	0.0082	n.s	n.s	*0.0459*	4.80	1.75-1.80
Age (in years)												
<25	-0.0046	n.s	n.s	-0.0276	n.s	n.s	-0.0988	n.s	n.s	-0.0679	n.s	n.s
25-50	-0.0211	n.s	n.s	0.0151	n.s	n.s	0.0166	n.s	n.s	0.0452	n.s	n.s
>50	0.0203	n.s	n.s	0.0262	n.s	n.s	**0.1000**	4.73	1.65-1.70	**0.1267**	7.40	1.75-1.80
Duration of unemployment												
< 13 weeks	*-0.0705*	4.82	1.55-1.60	-0.0477	n.s	n.s	-0.0847	n.s	n.s	0.0106	n.s	n.s
13-52 weeks	-0.0194	n.s	n.s	-0.0097	n.s	n.s	0.0082	n.s	n.s	0.0329	n.s	n.s
> 52 weeks	*0.0443*	6.06	1.65-1.70	**0.0503**	8.19	1.65-1.70	**0.0875**	8.24	1.75-1.80	**0.1125**	13.26	1.80-18.5
Without professional experience	-0.0364	n.s	n.s	-0.0040	n.s	n.s	-0.0781	n.s	n.s	-0.0703	n.s	n.s
Without professional training	**-0.0486**	8.44	1.40-1.45	-0.0046	n.s	n.s	-0.0089	n.s	n.s	0.0425	n.s	n.s
With high degree	*0.1354*	3.97	1.75-1.80	*0.1250*	3.54	1.75-1.80	0.0417	n.s	n.s	0.0083	n.s	n.s
With placement restrictions	-0.0215	n.s	n.s	0.0153	n.s	n.s	0.0513	n.s	n.s	0.1026	n.s	n.s
Rehabilitation attendant	-0.0400	n.s	n.s	0.0300	n.s	n.s	0.0000	n.s	n.s	0.0571	n.s	n.s
East Germany	Men						Women					
Total	**-0.0482**	31.23	1.15-1.20	**-0.0287**	12.11	1.25-1.30	**-0.0376**	33.02	1.25-1.30	*-0.0135*	12.11	1.25-1.30
Age (in years)												
<25	-0.0873	4.56	1.20-1.25	-0.0437	n.s.	n.s.	0.0069	n.s.	n.s.	0.0278	n.s.	n.s.
25-50	**-0.0439**	12.46	1.20-1.25	-0.0185	n.s.	n.s.	**-0.0420**	23.64	1.30-1.35	**-0.0219**	6.20	1.40-1.45
>50	**-0.0447**	16.11	1.00-1.05	-0.0130	n.s.	n.s.	-0.0142	n.s.	n.s.	-0.0020	n.s.	n.s.
Duration of unemployment												
< 13 weeks	**-0.1328**	24.79	1.00-1.05	**-0.1006**	15.25	1.05-1.10	**-0.1601**	30.64	1.00-1.05	**-0.0742**	7.56	1.15-1.20
13-52 weeks	*-0.0309*	5.16	1.25-1.30	-0.0163	n.s.	n.s.	**-0.0362**	11.01	1.25-1.30	-0.0076	n.s.	n.s.
> 52 weeks	-0.0009	n.s.	n.s.	-0.0018	n.s.	n.s.	0.0100	n.s.	n.s.	**0.0245**	9.74	1.60-1.65
Without professional experience	-0.0727	n.s.	n.s.	0.0069	n.s.	n.s.	-0.0143	n.s.	n.s.	0.0225	n.s.	n.s.
Without professional training	-0.0275	n.s.	n.s.	0.0120	n.s.	n.s.	**-0.0421**	10.20	1.10-1.15	-0.0215	n.s.	n.s.
With high degree	-0.0147	n.s.	n.s.	0.0074	n.s.	n.s.	*-0.1037*	5.95	1.00-1.05	*-0.0976*	5.08	1.00-1.05
With placement restrictions	0.0000	n.s.	n.s.	0.0189	n.s.	n.s.	*-0.0497*	4.11	1.00-1.05	-0.0166	n.s.	n.s.
Rehabilitation attendant	-0.0419	n.s.	n.s.	-0.0140	n.s.	n.s.	-0.0338	n.s.	n.s.	-0.0068	n.s.	n.s.

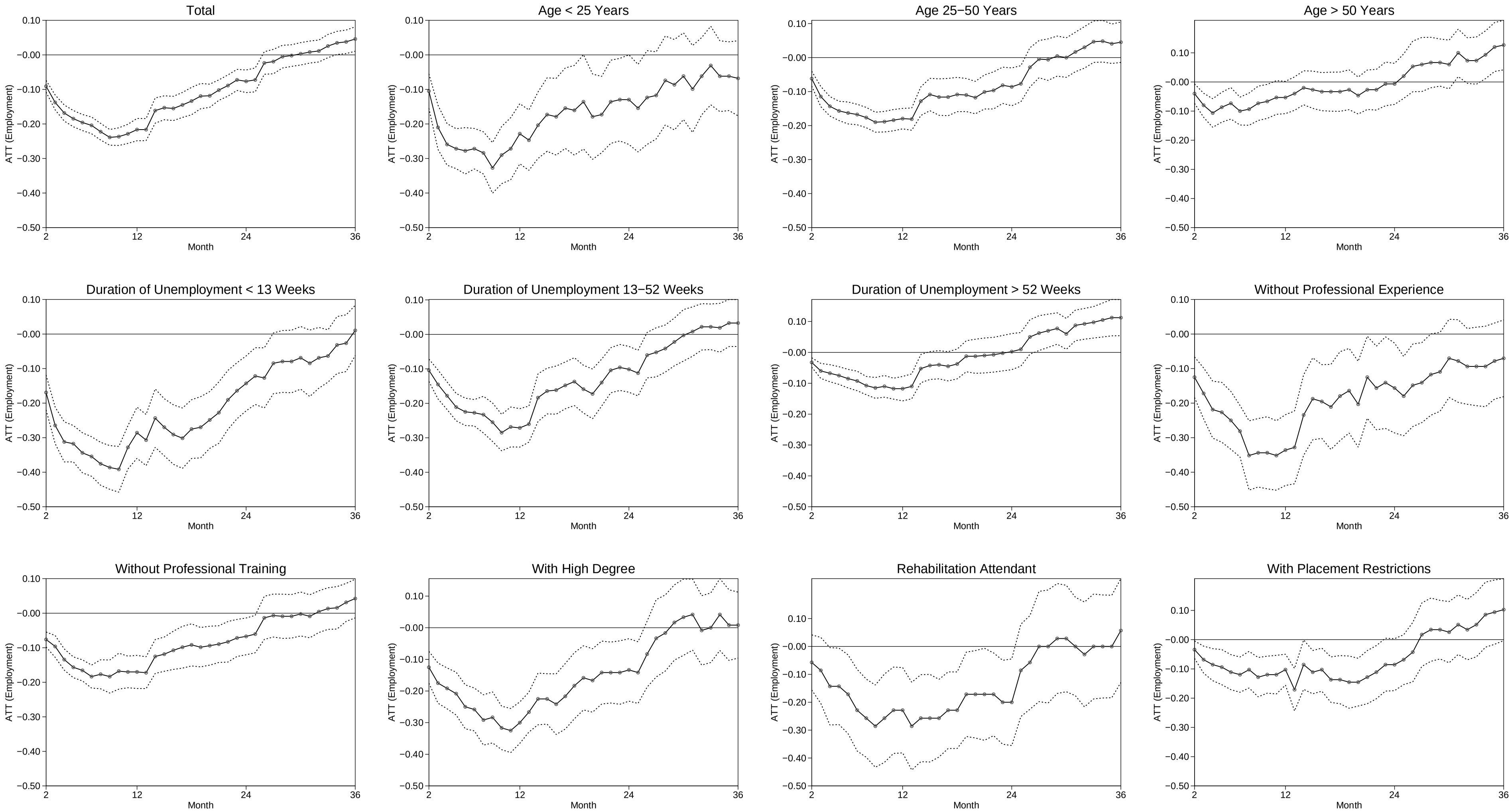
Bold letters indicate significance on a 1% level, *italic* letters refer to the 5% level.
n.s. = not significant

Fig. 1: Employment Effects for Men in West Germany (February 2000 - December 2002)



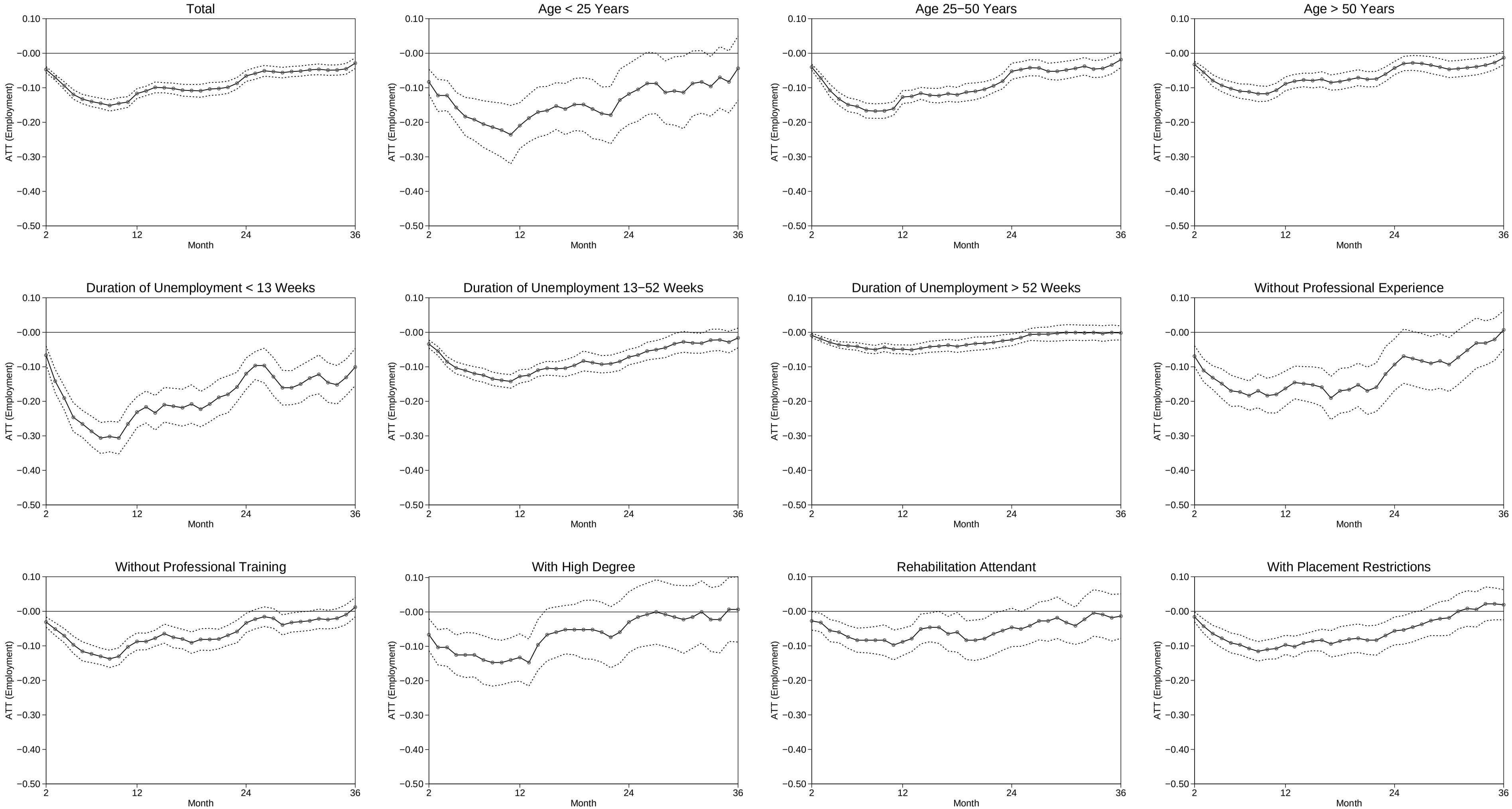
Solid line describes the monthly employment effect. Dotted lines are the upper and lower 95% confidence limits.
Month 2 refers to February 2000, month 12 = December 2000, month 24 = December 2001, month 36 = December 2002.

Fig. 2: Employment Effects for Women in West Germany (February 2000 - December 2002)



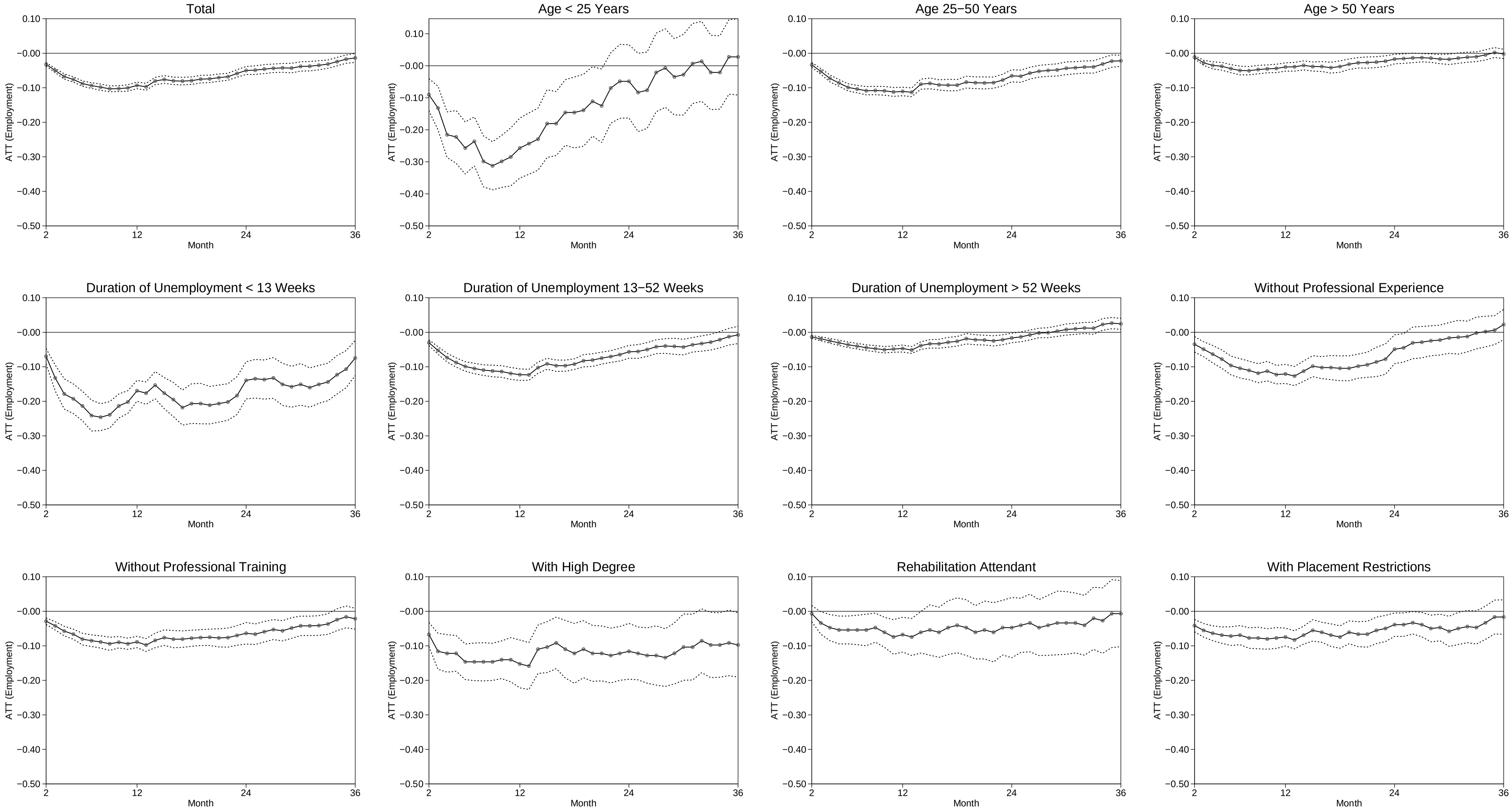
Solid line describes the monthly employment effect. Dotted lines are the upper and lower 95% confidence limits.
Month 2 refers to February 2000, month 12 = December 2000, month 24 = December 2001, month 36 = December 2002.

Fig. 3: Employment Effects for Men in East Germany (February 2000 - December 2002)



Solid line describes the monthly employment effect. Dotted lines are the upper and lower 95% confidence limits.
Month 2 refers to February 2000, month 12 = December 2000, month 24 = December 2001, month 36 = December 2002.

Fig. 4: Employment Effects for Women in East Germany (February 2000 - December 2002)



Solid line describes the monthly employment effect. Dotted lines are the upper and lower 95% confidence limits.
Month 2 refers to February 2000, month 12 = December 2000, month 24 = December 2001, month 36 = December 2002.