Disentangling Treatment Effects of Active Labor Market Policies: The Role of Employment Histories*

Jochen Kluve[a]

Hartmut Lehmann[b]

Christoph M. Schmidt[c]

This version: September 15th, 2004

- Preliminary -

Abstract. This paper estimates treatment effects of two active labor market policies – a training program and a wage subsidy scheme – on employment probabilities. The analysis is based on unique data from the 18th wave of the Polish Labor Force Survey that contains detailed and extensive individual labor force status histories. We discuss three stages of an exact covariate matching procedure adapted to the specific nature of the data. Our study confirms and reinforces a point raised in recent research (Heckman and Smith 1999, 2004), that pre-treatment labor force status dynamics play a decisive role in determining program participation – an aspect often ignored in evaluation research. We implement a conditional difference-in-differences estimator of treatment effects based on these individual trinomial sequences of pre-treatment labor market status. The estimator employs a "moving window" technique that nicely controls for changes in the macroeconomic environment over time. Our findings suggest that training raises employment probability, while the wage subsidy scheme seems to lead to a negative treatment effect for men. Furthermore, we find that stratification of the matched sample can add considerable insight regarding treatment effect heterogeneity.

Keywords: Active Labor Market Policy, exact matching, moving window, determinants of program participation.

JEL: C49, J68

---

* This paper is a completely and substantially revised version of the earlier paper "Disentangling treatment effects of Polish Active Labor Market Policies: Evidence from Matched Samples". We are grateful to Boris Augurzky, Patrick Puhani, participants of ESPE 2002, Bilbao, Spain, ESEM 2002, Venice, Italy, and EALE 2002, Paris, France, and an anonymous referee for valuable comments.
[a] RWI-Essen, and IZA Bonn.
[b] University of Bologna, CERT Edinburgh, and IZA Bonn.
[c] RWI-Essen, Ruhr-Universität Bochum, CEPR London, and IZA Bonn.
Correspondence: Jochen Kluve, RWI-Essen, Hohenzollernstr. 1-3, 45128 Essen, Germany, kluve@rwi-essen.de.

1. Introduction

Over the last decade there has been much interest by labor economists in the evaluation of so-called Active Labor Market Policy (ALMP), i.e. policy measures such as training programs, wage subsidy schemes, or direct job creation in the public sector. These measures, generally, aim at increasing the employment probability and/or the earnings performance of program participants. In the US, experiences with both the implementation and evaluation of such programs date back well into the 1960s (Heckman, LaLonde and Smith 1999). In Europe, where unemployment had remained comparatively low until the 1980s, or even until the 1990s in some countries, running such programs and evaluating them is a rather recent phenomenon. Nevertheless, most countries in Western Europe have now utilized active labor market measures for many years, and have done so with substantial financial input in terms of fraction of GDP spent on these measures (see e.g. OECD 2000). Also the evaluation practice, while still lagging behind the US "evaluation culture" to some extent, has attained increasing interest – and funding – by European policy makers, both in individual countries and from the European Commission. Kluve and Schmidt (2002) give a detailed account of the European experience with Active Labor Market Policies, embedded in the context of the European Employment Strategy, and contrast this experience with the evidence from the US.

After the breakdown of the socialist regimes and the beginning of the "transition process" around 1990, Eastern European countries were confronted with the task of redesigning their welfare system. Suddenly facing substantial open unemployment, schemes for passive and active support of unemployed individuals had to be set up from scratch. Frequently this led to transition countries' adoption of Western schemes, often without much knowledge about their efficiency. Poland, too, implemented a system of unemployment benefit support, accompanied by a set of Active Labor Market Policies. Specifically, Poland has been running the following programs for unemployed persons: a training program, "Intervention Works", i.e. a wage subsidy scheme, and "Public Works", i.e. direct employment in the public sector.

While the importance of such programs – as expressed in government spending as GDP share – has declined in Poland over recent years, a few studies of Polish labor market employment dynamics and Active Labor Market Policy evaluation were undertaken in the late 1990s (Góra and Schmidt 1998, Puhani 1998). In this paper, we will build on earlier work on program evaluation in Poland (Kluve, Lehmann and Schmidt 1999) and provide an in-depth investigation of the Polish experience with Active Labor Market Policy in the mid-1990s, specifically the years 1992 to 1996.

There is a set of features that make this study particularly interesting. First, we use data from the 18th wave of the Polish Labor Force Survey (PLFS). The data were collected in August 1996, and contained a supplementary questionnaire on past labor market experience of respondents. This supplement generated a unique set of individual employment histories dating from January 1992 until August 1996 and comprising a person's labor force status for every single month. The monthly labor force status captures employment, unemployment, inactivity, and participation in an active labor market program, as well as a set of other states, such as caring for a child etc., that are of minor interest to our study.

Second, the evaluation is set against the background of a country in the early years of transition. This implies a rapidly changing macroeconomic environment, making it indispensable to develop a treatment effect estimator that can account for these changes in an appropriate manner. Third, in addition to estimating treatment effects on the basis of individual employment histories, we can use the detailed monthly data to investigate further how important such labor market histories are in fact for determining participation in the program, and hence the evaluation approach. Recent research (Heckman and Smith 2004, complementing Heckman and Smith 1999) suggests that labor force status dynamics play the central role in driving participation dynamics. We will reinforce this point made by Heckman and Smith on the basis of a different data set, different active labor market programs implemented in a different country, in an entirely different context.

The core part of our analysis is the development of a matching estimator based on individual pre-treatment labor force status sequences. This creates a "moving window" structure that allows for individually flexible entry into and exit out of the program, hence conditioning on covariates and employment histories at exactly the month of program start, and comparing outcomes at exactly the month of program termination. Clearly, while increasing comparability of treated and comparison units, this procedure also nicely controls for changes in the macroeconomic environment. Our approach is delineated using three matched samples, for two active policy measures – training and intervention works - each. First, a "raw" sample (A) of program participants and a comparison group consisting of those untreated individuals that were unemployed at least once over the sample period. Second, a sample (B) where the comparison group is matched on a set of covariates, in particular taking into account the local labor market context, a variable whose importance in program evaluation is e.g. pointed out in Heckman, Ishimura and Todd (1997). Third, a sample (C) matched on both covariates and four-quarter individual pre-treatment employment histories, in the spirit of Card and Sullivan (1988).

The paper is organized as follows. Section 2 describes our data and the matching approach to program evaluation. In section 3 we discuss the matched samples, focusing on the timing of interventions and the role of pre-treatment labor force status histories. Section 4 presents our estimation strategy and estimation results. Section 5 concludes.

2. Data and Methods

2.1 The Data

We employ data from the 18th wave of the Polish Labour Force Survey (PLFS) as of August 1996. The PLFS is a quarterly rotating panel introduced in May 1992. The distinct feature of the August 1996 wave is a supplementary questionnaire containing retrospective questions on individual labor market behavior. Specifically, the questionnaire allows constructing individual employment histories on the basis of labor force status in every single month. Possible states are employed, unemployed, inactive, program participation, etc. (see below). The individual histories cover the 56-month-period from January 1992 to August 1996.

Our evaluation of the Training and Intervention Works programs is based on considering (a) pre-treatment labor force status information over a period of 4 quarters, i.e.12 months, and (b) post-treatment employment outcomes over a period of 3 quarters, i.e. 9 months. Given an overall sampling period from January 1992 until August 1996, we therefore focus on individuals whose treatment started after December 1992 and ended before December 1995. The analysis takes into account all individuals who experienced at least one spell of unemployment during the observation period. For both treated units and potential comparison units this ensures consideration of individuals potentially eligible for participation in ALMP measures offered by the employment offices. We discuss sample composition in more detail in section 3.1.

In order to be able to handle such rich data, we had to condense the information contained in the individual labor market histories. Monthly entries entail, for instance, states such as "employed", "unemployed", "receiving unemployment benefits", "maternal leave", etc. Furthermore, individual histories indicate whether and when an individual took part in an ALMP course. We compress the 30 possible monthly states occurring in the data into the three labor market states "employed" (henceforth denoted "1"), "unemployed" (denoted "2"), and "out-of-the-labor-force" (denoted "0"). Information on treatment participation is stored separately. Kluve et al. (1999) give a more detailed account of data transformation and adaptation. The resulting structure of individual spells for treatment and potential comparisons will be illustrated further in section 3.2.

In the estimation of individual treatment effects we consider two distinct measures of Polish ALMP, Training and Intervention Works[1]. Training is meant to enhance, or at least sustain,

[1] A third measure of Polish ALMP, Public Works (=direct job creation in the public sector), has been left out in this study for the sake of brevity, and due to very small sample sizes. Cf. also Kluve et al. (1999), Puhani (1998).

individual human capital during a period of unemployment. The Polish Training measure for the unemployed is training off-the-job whose final aim is raising the unemployed person's probability of re-employment in a regular job.

Wage subsidy schemes like the Polish Intervention Works also have a human capital enhancing or -preserving aspect. However, the enhancement or preservation of a person's human capital takes place on-the-job. This human capital component of the program is thought to increase the chances of a participant to find regular, non-subsidized employment at the same firm or elsewhere after the end of the program. In addition, if there is asymmetric information about the productivity of potential employees, wage subsidy schemes are designed to facilitate temporary job matches that might translate into regular and lasting matches at the same firm once the subsidy ends. A crucial feature of ALMP regulation in the reported period, however, was that participation in Intervention Works was considered by the law like any other employment spell, hence entitling individuals to a new round of benefit receipt, given the subsidized job lasted at least six months. Taking part in a Polish training measure for the unemployed, on the other hand, did not renew a person's benefit eligibility since this training was done off-the-job.

2.2 Matching Method

Program evaluation aims at estimating causal effects of treatments, i.e. changes in the outcome variable of interest that are due to participation in the treatment. The application of matching methods for treatment effect estimation has become quite popular over recent years, and several variants of matching estimators are now routinely applied.[2] The causal model underlying this approach has become known as the "Potential Outcome Model" and is based on work by Neyman (1923 [1990], 1935), Fisher (1935) and Rubin (1974, 1977; see also Holland 1986 and Kluve 2004 for discussion). The model formalizes the idea that, in order to infer a causal effect of the treatment on the outcome variable, it is necessary to identify the counterfactual, i.e. what would have happened to the treatment group if it had not been exposed to treatment? Then the causal effect of treatment is given by the difference between the factual (=exposed to treatment) and counterfactual (=not exposed to treatment) outcomes.

Let the binary variable $D \in \{0,1\}$ indicate the treatment received, i.e. $D = 1$ if the unemployed individual participates in the program. For each person we observe the treatment that

[2] Much research has been conducted in labor economics and econometrics on the practical and theoretical properties of matching estimators. See, for instance, the debate between Dehejia/Wahba and Smith/Todd (Dehejia and Wahba 1999, Smith and Todd 2004a, 2004b, Dehejia 2004) and a recent symposium in the Review of Economics and Statistics (2004, Vol. 86, No. 1, pp. 1-194).

she received, and the outcome associated with this treatment, i.e.

$$Y = Y_0 \quad if \;\; D = 0,$$
$$Y = Y_1 \quad if \;\; D = 1,$$

where the variable Y captures post-treatment outcomes of the variable of interest, i.e. individual labor market performance such as employment probability. Thus, the unit level causal effect given by $\Delta = Y_1 - Y_0$ is never directly observable. The essential conceptual point is that nonetheless each individual has two possible outcomes associated with herself, where one realization of the outcome variable can actually be observed for each individual, and the other one is a counterfactual outcome.

Since individual-level effects cannot be observed, the estimand of interest should be a measure that summarizes individual gains from treatment appropriately. One parameter that has received particular interest in the program evaluation literature is the average treatment effect for the treated population (ATET),

$$E(\Delta \mid D = 1) = E(Y_1 - Y_0 \mid D = 1) = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)\,,$$

where the expectations operator E(.) denotes population averages. The parameter is generally not identified from observational data: Whereas the first of the population averages in the ATET parameter can be identified for the treatment group subsample, the counterfactual expectation $E(Y_0 \mid D = 1)$ is not identifiable without invoking further assumptions, since the outcome under no-treatment is not observed for the treated population. This is precisely the counterfactual of interest: What outcome would the treated units have realized if they had not been exposed to the treatment?

If treatment is not randomly assigned, matching intends to mimic a randomized experiment ex post. This strategy is feasible if there is only "overt bias" (Rosenbaum 1995), i.e. treatment and comparison group differ prior to treatment only in observable variables that matter for the outcome under study. Let X denote the vector of observed pre-treatment variables, or covariates. Then the concept of "selection on observables" is formalized in the following identifying assumption: The assignment mechanism D is independent of the potential outcomes Y_0,Y_1 conditional on X (Rubin 1974, 1977). This assumption is commonly referred to as unconfoundedness (Imbens 2004). By the unconfoundedness assumption it is possible to replace the no-treatment outcome for the treated population with the no-treatment outcome of the non-treated, i.e. comparison, population:

$$\begin{aligned} E(\Delta \mid X, D = 1) &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1) \\ &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 0) \end{aligned}$$

This covariate-adjusted ATET is identified from observable data.

3. Analyzing Matched Samples

3.1 Composition of Matched Samples

For each of the two active labor market measures under scrutiny – Training and Intervention Works – we analyze treatment effects and illustrate the role of employment histories using three samples. We consider those observations in the PLFS that have at least one spell of unemployment over the sampling period January 1992 to August 1996. The three matched samples are then defined as follows.

> Sample A: A comparison unit is matched to a treated unit if his or her labor market history is observed without substantial gaps for 12 months preceding the start of treatment and for 9 months succeeding the end of treatment. The contents of the labor market history is not used, and no restrictions on covariates are imposed.
>
> Sample B: A comparison unit is matched to a treated unit if the requirement for sample (A) is met, and if he or she is identical in observable covariates age, gender, education, marital status, and region.
>
> Sample C: A comparison unit is matched to a treated unit if the requirements for sample (B) are met, and if he or she displays an identical 4-quarter (12-month) pre-treatment labor market history at the exact same point in time as the treated unit.[3]

The matching algorithm used to construct samples (A) through (C) applies exact covariate matching within calipers.[4] For all three samples, if a treated individual finds any matching partner among the potential comparisons, this observation is retained. The algorithm allows for an oversampling procedure, i.e. a treated unit may be assigned more than one comparison unit. The build-up from sample (A) to (C) reflects our conviction that timing is the pivotal aspect of comparison group construction in a transition economy.

The firmness in requirements (A) to (C) increases substantially. While under the weak precondition of Sample (A) no treated unit is lost in the matching process, and almost all potential comparisons are used, under requirement (C) some treated units do not find matching partners, and

[3] We consider 6 age categories, 3 education categories, gender, marital status, and 49 regions, resulting in 3528 different cells for sample (B). Including a 4-quarter sequence of a trinomial labor market outcome variable (cf. section 3.2) increases the number of cells to $3528*3^4$=285,768 cells for sample (C).

[4] See Augurzky and Kluve (2004) on the relative performance of different matching algorithms.

the number of matched comparison units is far smaller. Thus, algorithm (C) proceeds with replacement: some comparison units are matched to more than one treated individual. Samples (A) and (B) are constructed from potential comparison units with replacement, too, but here we use only the join of sets over matched comparison units.

Table 1 presents sample sizes and covariate means for the resulting samples. We observe that there is a reduction in the number of treated units who find matching partners from (A) to (C) of almost one third for Training, and almost one quarter for Intervention Works. Due to matching-with-replacement, samples (C) contain comparison units matched to more than one treated unit. With less than one percent, the number is very low for Training, and with approximately one tenth it is also fairly low for Intervention Works. Table 1 also shows that Training participants on average are better educated, somewhat younger and more likely to be female than Intervention Works participants. The distribution of the regional information that we use, i.e. the 49 Polish "voivodships", is presented in Figure 1 for sample A. Clearly, only sample (A) would display imbalances in the covariates, since matching in samples (B) and (C) by definition produces balance since it conditions on identicalness in observed characteristics age, education, sex, marital status, and local labor market (region).

For sample (B), this reflects a limited number of matching variables that are all categorical. Here, exact matching performs quite well: despite the substantial number of cells, approximately 9 out of 10 of treated units find a comparison unit. This number is further reduced in sample (C), when conditioning on identical 4-quarter pre-treatment histories. Given the strength of this restriction, however, the resulting number of matched treated units seems satisfactory.

3.2 The timing of interventions

In sample (C) we require treated and matched comparison units to display an identical pre-treatment history. To achieve comparability across the three samples (A) to (C), we impose the requirement on samples (A) and (B) that we observe any history at all in the year preceding treatment, although the precise information what history was experienced is not used in matching. Moreover, to allow an assessment of post-treatment labor market performance, we require treated units and comparison units in all samples to have a complete post-treatment sequence of labor force status in the nine months after treatment. Monthly employment information is condensed into a sequence of three quarters of a multinomial outcome variable (0,1,2) denoting labor force status (out-of-the-labor force, employed, unemployed).

For comparison units in sample (C) this procedure implies that they will only be matched to a treated unit if, in addition to being identical in the other covariates, they have an identical past 4-quarter employment history looking back from the point in time – the exact month – when the treated unit entered the program. Correspondingly, the 3-quarter outcome sequence for this matched comparison unit will be evaluated exactly congruent with the treated unit's 3-quarter post-treatment outcome sequence, i.e. after the treated person leaves the program. This approach accomplishes to define "treatment start" and "treatment stop" for comparison units, points in time that otherwise are not defined. Moreover, treated and untreated units are always compared during the same period, such that changes in general economic conditions, even on the local labor market level, are controlled for.[5]

Figure 2 illustrates the procedure for samples (A) and (B), in which the timing structure is considered, but the content of individual labor force status histories does not matter. Figure 3 delineates the approach for sample (C), where one or more controls are matched to a treated unit on the basis of identical pre-treatment employment histories at the same point in time. The figure also shows how the 12-month-sequence is condensed into the comparable 4-quarter-structure. Figures 2 and 3 show how this method generates a "moving window" as the algorithm advances through the spells of treated units one after the other searching for comparable untreated units at the corresponding points in time.

3.3 Pre-Treatment Histories

A central aspect of program evaluation regards the process that determines participation and non-participation in the program, and the potential problem of participants self-selecting into the treatment on the basis of observed or unobserved information. In the US, where ALMP measures were first evaluated, interest was mainly in the earnings performance of participants. In considering the determinants of participation, the focus then, logically, was on the differences in pre-treatment earnings performance of program participants and non-participants. In the context of such difference-in-differences estimation approaches, Ashenfelter (1978) already pointed to a potentially serious limitation of this procedure when he observed a relative decline in pre-treatment earnings for participants in subsidized training programs. This empirical regularity has been called "Ashenfelter's dip" and has been confirmed by subsequent analyses of many other training and adult education programs (cf. Bassi 1983, Ashenfelter and Card 1985, LaLonde 1986, Heckman, LaLonde, and Smith 1999). For instance, Ashenfelter and Card (1985) apply a model that focuses

[5] Such changes did indeed occur in Poland during transition. For instance, overall unemployment displayed an inverted U-shape over our sampling period, increasing from 13.3% (1992) via 14.0% (1993) to 14.4% (1994), and then falling again to 13.3% (1995) and 12.4% (1996).

on earnings changes as the determinants of participation. This line of thought was a logical consequence of Ashenfelter's discovery, and the main objective of the program, and resulted in analyses using earnings histories to eliminate differences between participants and nonparticipants. Clearly, the fact whether the pre-program earnings dip is transitory or permanent determines what would have happened to participants had they not participated, and the validity of any estimation approach depends on the relationship between earnings in the post-program period and the determinants of program participation (Heckman and Smith 1999).

This rather established observation that it is earnings dynamics that drive program participation has lately been put into serious question by Heckman and Smith (1999), who argue that it is rather labor force dynamics that determine participation in an ALMP program, a point they reinforce in their recent in-depth analysis of the determinants of program participation (Heckman and Smith 2004). This point had implicitly been made before by Card and Sullivan (1988), who analyze training effects conditional on pre-program employment histories. Furthermore, Heckman and Smith (1999) argue for a distinction between employment dynamics – indicating whether an individual is employed or not – and labor force dynamics, incorporating also whether a nonemployed person is either unemployed or out-of-the-labor-force. Their conclusion is "that labor force dynamics, rather than earnings or employment dynamics, drive the participation process" (Heckman and Smith 1999). Therefore, we extend the "employment history setting" considered in Card and Sullivan (1988) to a "labor force status history setting", reflecting also movements in and out of inactivity. This approach is delineated above in section 3.2.

Figures 4 and 5 draw the distributions of pre-treatment labor market histories for samples (A) and (B) for both Intervention Works (Fig.4) and Training (Fig.5)[6]. Representing a 12-month labor force status sequence with 4 quarterly realizations of a trinomial variable (0,1,2) yields 81 possible sequences ("0000" to "2222"). For the purpose of illustrating the distributions – and only for that purpose – we classify these 81 sequences into 11 categories (see Appendix A), so that on the abscissa the bottom categories contain "inactive" sequences (mostly '0's), the middle categories comprise "unemployed" sequences ('2's), and the top categories represent "employed" sequences ('1's). Categories 1, 6, and 11 exclusively embody the straight sequences (i.e. "0000", "2222", and "1111", respectively).

Thus, of the three peaks we observe in most of the graphs in Figures 4 and 5, the left peak represents "inactive" histories, because histories with a low order number contain many '0's. Accordingly, the peak in the middle expresses "unemployed" histories, and the peak to the right

[6] Clearly, in sample (C) these distributions will be balanced.

depicts "employed" histories. In terms of balancing of distributions, the picture is almost the same for Figures 4 and 5. Both samples (A) and (B) display only limited accordance in pre-treatment histories for treated and comparison units. The figures also show that treatment individuals in Training are quite different from those in Intervention Works. For the Training participants, the fractions of "employed" and "unemployed" histories are quite close to each other, while in the Intervention Works sample we observe a far larger fraction of "unemployed" histories among the treated. Moreover, for both Training and Intervention Works the comparison samples (A) and (B) are too "successful" in that they contain too many "employed" sequences relative to "unemployed" sequences in order to be comparable to the treated units, where "unemployed" sequences dominate.

It is interesting to note that the comparison group in sample (B) should have improved on the comparison group in sample (A), since sample (B) is matched on covariates age, education, sex, marital status, and region, but there is very little difference in the distribution of employment histories moving from (A) to (B), for both Training and Intervention Works. Finally, note that there seem to be only few "weird" histories, i.e. histories in which individual constantly change labor force status. This is especially true for the Intervention Works samples, where the majority of treated units, by far, has been unemployed for the full 4 quarters preceding treatment.

4. Empirical results

4.1 Distributions of outcomes

Figures 6 and 7 plot distributions for the post-treatment employment success for treated units and comparison units in samples (A) to (C). There are 27 possible labor market status sequences capturing employment performance in the three quarters succeeding treatment (cf. also Figures 2 and 3). Similar to our presentation of pre-treatment labor market histories, we classify these 27 possible sequences of 3 quarterly realizations of a trinomial variable into 9 categories for illustration purposes. This categorization is outlined in Appendix A. Once more, bottom categories contain "inactive" sequences (category 1="000"), middle categories include "unemployed" sequences (category 5="222"), and top categories comprise "employed" histories (category 9="111"). Accordingly, in the graphs the left peak depicts "inactive" sequences, the middle peak "unemployed" sequences, and the right peak represents "employed" histories.

Looking at the Intervention Works samples in Figure 6, we find that in all samples the "unemployed" sequences are clearly predominant for the treated units. At the same time, comparison units display rather successful labor market histories in samples (A) and (B). In sample (C) this picture changes considerably, and a larger fraction of comparison units also displays "unemployed" histories. However, the comparison group still fares visibly better than the program participants. Sample (C) therefore indicates that during the 9 months directly succeeding participation in Intervention Works the treated units seem to be on average marginally – possibly insignificantly – less successful in finding employment than the comparison units.

For the Training samples shown in Figure 7 we find slightly different results. Similar to what we have seen for the pre-treatment sequences of these samples (Figure 5), the "employed" and "unemployed" peaks have more or less the same height also for the post-treatment sequence. But while for samples (A) and (B) the "employed" peak is higher for comparison units than for treated units, and the "unemployed" peak is higher for treated units than for comparison units, this relation switches for sample (C). In (C) treated units display on average a slightly more successful post-treatment labor market sequence than corresponding comparisons. This would be an indication of a slightly – possibly insignificant – positive treatment effect of Training.

Taken together, Figures 6 and 7 display four important patterns. First, moving from (A) to (C) we do not observe much variation in the distributions for treated units. Thus, the fact that we lose some treated units while increasing matching requirements does not seem to play an important role. Second, the distributions for the comparison groups do not change much when moving from (A) to (B) taking covariates, but not the employment histories, into account. Third, without conditioning on pre-treatment labor market histories the comparison samples apparently contain too many "successful" individuals – a pattern which we already observed for pre-treatment labor force status sequences in Figures 4 and 5. For samples (A) and (B) this would probably result in too negative an estimate of treatment effects. Fourth, across comparison units and treated units we observe clearly more "successful" outcomes for Training than for Intervention Works. This, too, is not surprising, as we noticed a similar relation for pre-treatment labor market history distributions (Fig. 4 and 5).

4.2 Treatment effect estimation

Our aim is to identify treatment effects of two different measures of Polish active labor market policy, Intervention Works and Training, which we consider separately in the empirical analysis. For purposes of the formal exposition of our estimation approach we consider a single generic intervention. Furthermore, we explicitly require that treated units be matched with comparison units from the identical set of observed pre-treatment and post-treatment months. Any reference to the time period is therefore omitted from the formal exposition as well.

In addition to the terminology introduced in section 2, let N_1 denote the number of treated units, with indices $i \in I_1$, and N_0 the number of potential comparison units, with indices $i \in I_0$. Potential labor market outcomes in post-treatment quarter q $(q = 1, 2, 3)$ are denoted by Y_{qi}^1, if individual i received treatment, and by Y_{qi}^0, if individual i did not receive treatment. Outcomes are defined as multinomials with three possible realizations ('0'=out-of-the-labor-force, '1'=employed, '2'=unemployed), extending the formulations of Card and Sullivan (1988) from a binomial to a trinomial setting.

We can only observe one of the two potential outcomes Y_{qi}^1 and Y_{qi}^0 for a given individual. This actual outcome is denoted by Y_{qi}. The objective is then to formally construct an estimator of the mean of the unobservable counterfactual outcome $E(Y_{qi}^0 \mid D_i=1)$. Following the quarterly sequence of labor market outcomes might be too detailed, though, for a direct economic interpretation of results. Thus, to condense the available information further, the post-intervention labor market success of each individual i is summarized by the individual's average employment rate over the three quarters following the intervention. Using indicator function $\mathbf{1}(.)$, these employment rate outcomes are $\frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=1)$.[7] *Observed* outcomes for individual i can then be written as

$$(4) \qquad \frac{1}{3}\sum\nolimits_q \mathbf{1}(Y_{qi}=1) \quad = \quad \frac{1}{3}\left(D_i \sum\nolimits_q \mathbf{1}(Y_{qi}^1=1) + (1-D_i)\sum\nolimits_q \mathbf{1}(Y_{qi}^0=1)\right) \quad ,$$

and the impact of the intervention on the average labor market status of individual i can be expressed as

$$(5) \qquad \Delta_i = \frac{1}{3}\left(\sum\nolimits_q \mathbf{1}(Y_{qi}^1=1) - \sum\nolimits_q \mathbf{1}(Y_{qi}^0=1)\right)$$

for average employment rates. The parameters of interest in our evaluation analysis are weighted population averages over these individual treatment effects, the mean effect of treatment on the

[7] Kluve et al. (1999) extend this setting to considering both employment and unemployment rates, so that corresponding outcomes would be $\frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=w)$, where $w \in \{1,2\}$. Comparing employment and unemployment rate treatment effects shows for instance that exits to inactivity play a much larger role for women than for men. Moreover, Kluve et al. (1999) also consider the medium run, i.e. 6 post-treatment quarters, while we focus on the short-term case here. The extension to any number of post-treatment periods is straightforward.

treated for types of individuals characterized simultaneously by specific sets of characteristics X; and labor market histories before treatment h_i,

$$(6) \qquad E(\Delta_i \big| X_i, h_i, D_i = 1) \quad = \quad E\left(\frac{1}{3}\left(\sum_q \mathbf{1}(Y_{qi}^1 = 1) - \sum_q \mathbf{1}(Y_{qi}^0 = 1)\right) \big| X_i, h_i, D_i = 1\right) \quad .$$

The less inclusive the chosen set of characteristics conditioned upon – i.e. the more specific characteristics are included in X – the larger is the population of treated individuals over which the conditional mean is taken. As laid out above, previous labor market histories h_i are captured by the sequence of labor market states in the four quarters preceding the intervention.

Our approach to combine the population averages of the treatment effects for individuals in a given history-specific "cell" – characterized by demographic and other characteristics, in particular labor market history – gives us considerable flexibility in addressing the economic interpretation of results. The standard approach to evaluation would be to consider the distinction of type-history cells primarily as a device to achieve comparability of treatment and comparison units (see below). The ultimate interest there typically lies in the average treatment effects over the joint support of X and h given $D=1$,

$$(7) \qquad M = \sum_s w_s E(\Delta \mid s, D = 1) \,,$$

with s indicating any possible combination of X and h, and w_s representing the corresponding relative frequency in the treatment sample. By contrast to this standard approach, in what follows we will consider appropriate subsets of this joint support.

How does our particular observational approach – matching – facilitate the estimation of these parameters of interest? In randomized experiments the counterfactual expected values under no intervention can simply be estimated for intervention recipients by the mean values of the outcome for randomized-out would-be recipients. As we have shown in section 2, matching methods can recover the desired counterfactual for a nonexperimental comparison group: Within each matched set of individuals, one can estimate the treatment impact on individual i by the difference over sample means, and one can construct an estimate of the overall impact by forming a weighted average over these individual estimates.

Matching estimators thereby approximate the virtues of randomization mainly by balancing the distribution of observed attributes across treatment and comparison groups, both by ensuring a common region of support for individuals in the intervention sample and their matched comparisons and by re-weighting the distribution over the common region of support. The central identification assumption is that of mean independence of the labor market status Y_{qi}^{0} and of the treatment indicator D_i, given individual observable characteristics. In our specific application these conditioning characteristics are the demographic and regional variables X_i and the pre-treatment history h_i, i.e. from equation (2) in our case,

$$(8) \qquad E(\mathbf{1}(Y_{qi}^{0}=1) \mid X_i, h_i, D_i=1) \quad = \quad E(\mathbf{1}(Y_{qi}^{0}=1) \mid X_i, h_i, D_i=0) \quad .$$

Thus, by conditioning on previous labor market history we exploit the longitudinal nature of our data.

In a standard difference-in-differences approach pre-treatment and post-treatment outcomes are typically treated symmetrically; the identifying assumption is that the change in outcomes that treated individuals would have experienced had they not received treatment, would have been the same change – on average – that untreated individuals experience during the same period. This assumption accounts for the phenomenon that treatment units typically experience lower pre-treatment outcomes, even though they might be otherwise identical to comparison units. It does not lend itself naturally to the analysis of categorical outcome variables, though. In this context, a natural generalization of the difference-in-differences idea is to condition on the specific realization of the outcome variable in the pre-treatment period, as we do here. This is possible, since due to the categorical nature of the outcome the conditioning remains tractable. Card and Sullivan (1988) and Heckman et al. (1997) advocate such difference-in-differences approaches (cf. also Schmidt 1999).

Our matching estimator is one of oversampling exact covariate matching within calipers, allowing for matching-with-replacement. Our particular attention to pre-treatment labor market histories implements this idea of a generalized difference-in-differences juxtaposition between treated units and comparison units. Due to the relevance of the previous history for subsequent labor market success – state dependence is one of the issues most discussed in the labor literature – we also emphasize this variable in the construction of the estimates. Specifically, for any treatment history h for which at least one match could be found, we estimate the impact of the intervention by

$$(9) \qquad \hat{M}_h \quad = \quad \frac{1}{N_{1h}} \sum_{i \in I_{1h}} \left[\frac{1}{3} \sum\nolimits_q \mathbf{1}(Y_{qi}^{1}=1) \quad - \sum_{j \in I_{0h} \mid X_j \in C(X_i)} \frac{1}{n_{i0}} \left(\frac{1}{3} \sum\nolimits_q \mathbf{1}(Y_{qj}^{0}=1) \right) \right] \quad ,$$

where N_{1h} is the number of individuals with history h who receive the intervention ($N_1 = \sum_h N_{1h}$), I_{1h} is the set of indices for these individuals, $C(X_i)$ defines the caliper for individual i's characteristics X_i, and n_{i0} is the number of comparisons with history h who are falling within this caliper, with the set of indices for comparison-individuals with history h being I_{0h}. The standard error of the estimated treatment effect is then constructed as a function of the underlying multinomial probabilities. This procedure is outlined in Appendix B.

The overall effect of the intervention is estimated in a last step by calculating a weighted average over the history-specific intervention effects,

$$(10) \qquad \hat{M} = \sum_h \left[\frac{N_{1h}}{\sum_h N_{1h}} \hat{M}_h \right] ,$$

using the treated units' sample fractions as weights. The variance is derived as the corresponding weighted average of the history-specific variances.

4.3 Treatment effect results

Table 2 presents average treatment effects on the post-intervention employment rate for Intervention Works sample (C). The table shows how the overall treatment effect (-.126) is calculated by computing history-specific effects first. As explained above, for each treated unit, if he or she has more than one matched comparison unit, the comparison units' employment rates are averaged and handled as if they were the employment rate of only a single matched unit. The total effect is the weighted average of the history-specific effects using the treated units' sample fractions as weights.

Besides treatment effect calculation Table 2 shows the frequency with which labor market state sequences occurred in the data, thus picking up the theme of figure 4. We observe the same predominance of "unemployed" histories. The total treatment effect casts a rather negative picture on the Intervention Works program, suggesting that participation tends to lower post-treatment employment prospects. This is the treatment effect we believe to be the most credible, as it controls for pre-treatment employment experience of program participants. As discussed in extenso above, matching on the covariates does not achieve satisfactory balance of the employment history distribution.

Table 3 first reports the treatment effect estimates we would have obtained on the basis of samples (A) and (B). Unsurprisingly, in the light of our discussion of Figures 4 and 6, both estimates are similar in magnitude, and much more negative than the estimate derived from sample (C). In Table 3 we proceed to investigate the heterogeneity of the total treatment effect for sample C, by stratifying the sample along the lines of several characteristics. Specifically, we look at gender, at the date of program entry, early vs. late in the sampling period, and at specific labor market histories.

The simple stratification by gender reveals an interesting finding: The significantly negative full sample effect consists of a – more or less – zero treatment effect for women and a considerably larger negative effect for men. On the other hand, a subdivision by date of program entry that parts the observation period into two halves does not reveal any apparent influence of changes in the macroeconomic environment, or changes in the program implementation. The next step is to further refine cells and stratify the sample by both gender and date of program entry. These subsamples indicate that post-treatment employment prospects for male Intervention Works participants were quite unfavorable in the second period after July 1994, but particularly severe during the first period until June 1994. For women, effects for both time periods are insignificant.

Classification by labor market history allows us to look at the two major labor force status sequences that drive the peaks observed in Figure 4. For "employed" (1111) histories subsample sizes are rather small and the effects not well defined. For the subsample of "unemployed" (2222) histories, which entails almost 80% of total treated and comparison units, we find a significantly negative treatment effect close to the full sample effect. This is certainly no surprise, as the estimate of the full sample effect is dominated by the "2222" subsample effect. If we further stratify by labor market history and gender, treatment effects for the "1111" subsample remain insignificant for both men and women, while the "2222" subsample displays the same substantial male/female difference in the treatment effect that we have seen for the full sample.

Table 4 reports the same comparison between samples and various stratifications for Training. Both treatment effect estimates for samples (A) and (B) are insignificant, while the estimate obtained from sample (C) indicates that Training raises the individual employment probability by 13.8%. With regard to heterogeneous effects, a classification by gender does not seem to add any insights to the interpretation: Treatment effects for men and women are at best marginally significant, and would then be almost identical. Stratification by time period shows that the positive treatment effect occurred mostly before July 1994. However, it seems that in the first period mostly women benefited from training, whereas men benefited in the second period. Looking at a classification by labor market history, once more we find the "peaks" from Figure 5, indicating that the share of "1111" sequences is almost as large as the share of "2222" sequences. Subsample sizes here are too small to draw any conclusions.

From these calculations results the observation that an appropriate stratification of a matched sample can substantially contribute to disentangling and identifying heterogeneous treatment effects. In particular, the example of a simple classification by gender for the Intervention Works sample is striking: The overall negative effect is almost exclusively due to the dismal post-treatment labor market performance of male participants, in particular those who had been long-term unemployed prior to treatment.

In finding reasons for these negative treatment effects, it is sometimes suggested that it is stigma that causes participants of an employment program like Intervention Works to perform worse in the labor market than non-participants. Prospective employers identify participants as "low productivity workers" and are not willing to accept them into regular jobs. Another explanation, which might have particular merit in the Polish case, is benefit churning. Workers with long unemployment spells who have difficulty finding regular employment are identified by labor bureau officials and might only be chosen for participation in an employment scheme so that they re-qualify for another round of benefit payment.

While the presented evidence cannot pinpoint precisely the cause underlying the poor labor market performance of males participating in Intervention Works, stigmatization seems to be a less likely cause. For if participation in the scheme was a bad signal to prospective employers, it is not clear why this would not be the case for female participants. It may be that those males – males are for the most part heads of households – are targeted by labor bureau officials who have especially poor prospects for regular employment. Once the publicly subsidized job comes to an end, so officials might reason, they at least qualify for another round of unemployment benefits, if they cannot find regular employment elsewhere or if their subsidized job is not transformed into a regular job. It is probably not a mere coincidence that the large majority of Intervention Works jobs lasts six months, the length of time one needs to work within the year preceding benefit receipt in order to qualify for unemployment benefits.

In Table 5 we have a closer look at this idea that participation in Intervention Works might primarily be a vehicle to renew eligibility for unemployment benefits. The table shows the incidence of unemployment benefit receipt for men and women before and after the treatment for sample (C). The top panel indicates benefit receipt in at least two of the three months directly *preceding* treatment. The middle panel shows benefit receipt in at least two of the three months directly *succeeding* treatment. The bottom panel indicates benefit receipt in at least two months of each of the three quarters succeeding treatment, i.e. at least 6 out of 9 months. We observe that a substantial fraction of both treated and comparison units received pre-treatment benefits, although benefits do seem to play a more important role for treated units. This pattern is more pronounced for men. In the middle and bottom panel this situation aggravates substantially. While both short-term and medium-term benefit receipt played a minor role for comparison units, we observe that approximately 60% of the treated males received unemployment benefits in the quarter directly following treatment, and that more than half of the treated males received benefits during the whole 9-month post-treatment period. For females, this pattern is not quite as severe, but still post-treatment benefit receipt plays a major role for Intervention Works participants.

In addition to displaying treatment effects by sample and stratification, Table 6 presents treatment effect estimates for samples (C) obtained from a "counterfactual experiment". The first line reports the factual Intervention Works treatment effect estimate computed as shown in Table 2. This estimate tries to answer the question: "How much did Intervention Works participants benefit from participating in Intervention Works?" The second line reports a "counterfactual" Intervention Works treatment effect for Training participants, i.e. it tries to answer the question: "How much would Training participants have benefited, if they had participated in Intervention Works?" The estimate is obtained by history-wise reweighting the Intervention Works sample using the fraction of the treated units in the Training sample as weights. Looking at Table 2 this is the same as if for each history the second column contained the corresponding number of observations from the Training sample. Apparently, this reweighting by labor market history implicitly assumes that there are no relevant changes in other elements of *X*.

The estimate in the second row of Table 6 shows that, while the Intervention Works effect on Training participants still displays a negative sign, the effect is insignificant, so that Training participants participating in Intervention Works would have done better than Intervention Works participants themselves. Looking at the effects of Training on Training participants and Intervention Works participants, respectively, we find the counterpart to this result: Intervention Works participants participating in Training instead would have not gained as much from the treatment as Training participants themselves. Thus, persons with better observable and unobservable characteristics seem to have been targeted for the Training program.

The last two lines in Table 5 report differential treatment effects of Intervention Works vs. Training. The estimates represent the difference between the difference of treated and comparison units in Intervention Works (second to last column, Table 2) and the difference of treated and comparison units in Training. Once more, differences are taken history-wise and weighted using either Intervention Works participants or Training participants sample weights. Both estimates clearly show that Training is the superior ALMP relative to Intervention Works.

The methodology used in our paper allows us to evaluate ALMP at the individual level. It thus tells us that those persons participating in Polish Training programs have better employment prospects than they would have had had they not participated and also that they have better employment prospects than those who take part in Intervention Works. The methodology does not address the issue whether Training improves the overall performance of the labor market, i.e., for example, whether it lowers the aggregate unemployment rate. Even if Training is beneficial at the individual level, substitution effects - Training participants just "jump the queue" of those in line for regular jobs - could neutralize its impact at the aggregate level. On the other hand, the finding that a program is not even effective at the individual level, like the Polish Intervention Works scheme, helps us to focus attention on targeting issues and/or wrong incentive structures that distort the behavior of labor bureau officials and of the unemployed.

5. Conclusion

Over the last decade, there has been much interest by labor economists and econometricians – from both practical and theoretical perspectives – in the evaluation of treatments, in particular labor market interventions such as training and wage subsidy schemes. Building on a rather established culture of implementing and evaluating such Active Labor Market Policies in the US, Western European countries have caught up substantially in this regard. The next to follow are countries of Eastern Europe, often called "transition countries", that in the early years of transition in many cases implemented unemployment benefit schemes and active labor market programs similar to Western welfare systems. In the case of Poland, even though programs had to be built from scratch, already in the year 1996 data were collected that allow comprehensive evaluation of the country's ALMP measures during the first half of the 1990s.

In this study, we present an evaluation of two Polish active labor policies, Training and Intervention Works, i.e. a wage subsidy scheme. There is a set of features that we think make this study especially interesting. First, we use data from the 18th wave of the Polish Labor Force Survey that contain extensive and detailed information on individual labor force status histories from January 1992 until August 1996. Second, the evaluation is set against the background of a country in the early years of transition, a fact that implies a rapidly changing macroeconomic environment, making it indispensable to develop a treatment effect estimator that can account for these changes appropriately. Third, in addition to estimating treatment effects on the basis of individual employment histories, we use the detailed monthly data to investigate further how important such labor market histories are in fact for determining participation in the program.

The core part of the analysis is the development of a matching estimator based on individual pre-treatment labor force status sequences. This creates a "moving window" structure that allows for individually flexible entry into and exit out of the program, hence conditioning on covariates and employment histories at exactly the month of program start, and comparing outcomes at exactly the month of program termination. Clearly, while increasing comparability of treated and comparison units, this procedure also nicely controls for changes in the macroeconomic environment. Our approach is delineated using three matched samples (A) through (C) that differ in matching requirements.

We find that pre-treatment labor force status sequences contain indispensable information regarding selection into treatment and that controlling for these histories can eliminate a large part of the overt bias between treated and comparison units. This result confirms and reinforces the point made by Heckman and Smith (1999, 2004). Somewhat surprisingly, matching on a set of other covariates – including detailed regional information – does not seem to contribute much to improving the composition of the matched comparison group. In terms of treatment effect estimates we find a positive effect of training on the treated population, and a negative effect of Intervention Works. The latter is driven by strongly negative effects for men, which we attribute to a practice of "benefit churning", according to which men who were no longer eligible for unemployment benefit receipt participate in the program only so that this eligibility will be restored.

References

Ashenfelter, O. (1978) "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics* 60, 47-57.

Ashenfelter, O. and D. Card (1985) "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics* 67, 648-660.

Augurzky, B. and J. Kluve (2004), "Assessing the performance of matching algorithms when selection into treatment is strong" *IZA Discussion Paper* 1301, IZA Bonn.

Bassi, L.J. (1983) "The Effect of CETA on the Post-Program Earnings of Participants", *The Journal of Human Resources* 18, 539-556.

Card, D. and D. Sullivan (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica* 56, 497-530.

Dehejia, R. and S. Wahba (1999), 'Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs', *Journal of the American Statistical Association*, 94, 1053-1062.

Dehejia, R. and S. Wahba (2002), 'Propensity Score-Matching Methods for Nonexperimental Causal Studies', *The Review of Economics and Statistics*, 84, 151-161.

Dehejia, R. (2004), 'Practical Propensity Score Matching: A Reply to Smith and Todd', *Journal of Econometrics*, forthcoming.

Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Góra, M. and C.M. Schmidt (1998), "Long-term unemployment, unemployment benefits and social assistance: The Polish experience", *Empirical Economics* 23, 55-85.

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

Heckman, J.J., R.J. LaLonde and J.A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in: Ashenfelter, O. and D. Card (eds.): *Handbook of Labor Economics*, vol. III, Amsterdam et al.: North-Holland.

Heckman, J.J., and J.A. Smith (1999), "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies", *The Economic Journal* 109, 313-348.

Heckman J.J. and J.A. Smith (2004), "The Determinants of Participation in a Social Program: Evidence from a Prototypical Job Training Program", *Journal of Labor Economics* 22, 243-298.

Holland, P.W. (1986), 'Statistics and Causal Inference (with discussion)', *Journal of the American Statistical Association,* 81, 945-970.

Imbens, G.W. (2004), 'Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review', *Review of Economic and Statistics*, 86, 4-29.

Kluve, J. (2004), 'On the Role of Counterfactual in Inferring Causal Effects', *Foundations of Science*, 9, 65-101.

Kluve, J., H. Lehmann, and C.M. Schmidt (1999), "Active Labor Market Policies in Poland: Human Capital Enhancement, Stigmatization, or Benefit Churning?", *Journal of Comparative Economics* 27, 61-89.

Kluve, J. und C.M. Schmidt (2002), "Can training and employment subsidies combat European unemployment?", *Economic Policy* 35, 409-448.

LaLonde, R.J. (1986), "Evaluating the econometric evaluations of training programs with experimental data", *American Economic Review* 76, 604-620.

Neyman, J. (1923 [1990]), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.", translated and edited by D.M. Sabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923), 1-51 (Annals of Agriculture), *Statistical Science* 5, 465-472.

Neyman, J. (1935), with co-operation by K. Iwaszkiewicz, and S. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation", (with discussion), *Supplement to 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", *ZEW Disc. Paper 98-25*, Mannheim.

Rosenbaum, P.R. (1995), "*Observational Studies*", New York: Springer Series in Statistics.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66, 688-701.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics* 2, 1-26.

Smith, J.A. and P.E. Todd (2004a), 'Does Matching overcome LaLonde's critique of Nonexperimental estimators?', *Journal of Econometrics*, forthcoming.

Smith, J.A. and P.E. Todd (2004b), 'Rejoinder', *Journal of Econometrics*, forthcoming.

Appendix A. Categorizing labor market status sequences

Pre-treatment

Category	1	2	3	4	5	6	7	8	9	10	11
Histories	0000	0001	0012	0022	2201	2222	2220	2211	1102	1110	1111
		0010	0102	0202	2021		2202	2121	1012	1101	
		0100	1002	2002	0221		2022	1221	0112	1011	
		1000	0120	0220	2210		0222	2112	1120	0111	
		0002	1020	2020	2012		2221	1212	1021	1112	
		0020	0021	2200	0212		2212	1122	0121	1121	
		0200	1200		2120		2122		1210	1211	
		2000	0201		2102		1222		1201	2111	
			0210		0122				0211		
			2100		1220				2110		
			2010		1202				2101		
			2001		1022				2011		
			0110						0011		
			1010						0101		
			1100						1001		

Post-treatment

Category	1	2	3	4	5	6	7	8	9
Histories	000	001	210	220	222	221	012	110	111
		010	120	202		212	021	101	
		100	102	022		122	201	011	
		002						112	
		020						121	
		200						211	

Appendix B. Calculation of treatment effects and variances

The history-specific treatment effect estimator (9) is based on the differences in average employment rate outcomes between treatment and comparison units. One notable element of this estimator is that multiple comparison units matched to a single treated unit (due to the oversampling algorithm) are handled as if they were one single comparison unit. The variance for (9) is then composed of the sum of independent single variances of each of the employment rate averages entering (9) for "individual" treated and comparison units. This appendix illustrates the generic calculation of this individual variance, and how this yields variances for (9) and (10).

Within each stratum – defined by pre-treatment labor market history – employment success in the three post-treatment quarters is summarized by the average employment rate $\frac{\sum 1}{3}$. For the unrestricted multinomial model each of the $3^3=27$ possible outcomes is associated with a separate probability. For instance, conditional on the *k-th* history the probability to be employed in all subsequent quarters is $p(111 | h_k)$, the probability to be employed in the first and unemployed in the following two quarters is $p(122 | h_k)$, the probability to be unemployed in the first two and out-of-the-labor-force in the third quarter is $p(220 | h_k)$ etc. Let us order the 27 probabilities in the following way

$\frac{\sum 1}{3}=0$	$\frac{\sum 1}{3}=\frac{1}{3}$	$\frac{\sum 1}{3}=\frac{2}{3}$	$\frac{\sum 1}{3}=1$
$p(000\|h_k)=p_1$	$p(001\|h_k)=p_9$	$p(011\|h_k)=p_{21}$	$p(111\|h_k)=p_{27}$
$p(002\|h_k)=p_2$	$p(021\|h_k)=p_{10}$	$p(211\|h_k)=p_{22}$	
$p(020\|h_k)=p_3$	$p(201\|h_k)=p_{11}$	$p(101\|h_k)=p_{23}$	
$p(200\|h_k)=p_4$	$p(221\|h_k)=p_{12}$	$p(121\|h_k)=p_{24}$	
$p(022\|h_k)=p_5$	$p(010\|h_k)=p_{13}$	$p(110\|h_k)=p_{25}$	
$p(202\|h_k)=p_6$	$p(012\|h_k)=p_{14}$	$p(112\|h_k)=p_{26}$	
$p(220\|h_k)=p_7$	$p(210\|h_k)=p_{15}$		
$p(222\|h_k)=p_8$	$p(212\|h_k)=p_{16}$		
	$p(100\|h_k)=p_{17}$		
	$p(102\|h_k)=p_{18}$		
	$p(120\|h_k)=p_{19}$		
	$p(122\|h_k)=p_{20}$		

where $p_{27}=1-\sum_{m=1}^{26} p_m$. Then, for each individual i with history k (suppressing the subscripts h_k

for notational convenience)

(B1)
$$
\begin{aligned}
E(\frac{\sum 1}{3}) &= E[\frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=1)] \\
&= 0(p_1 + \ldots + p_8) + \frac{1}{3}(p_9 + \ldots + p_{20}) + \frac{2}{3}(p_{21} + \ldots + p_{26}) + 1p_{27} \\
&= \frac{1}{3}(p_9 + \ldots + p_{20}) + \frac{2}{3}(p_{21} + \ldots + p_{26}) + (1 - \textstyle\sum_{m=1}^{26} p_m) \\
&= \mu
\end{aligned}
$$

and

(B2)
$$
\begin{aligned}
Var(\frac{\sum 1}{3}) &= (-\mu)^2(p_1 + \ldots + p_8) + (\frac{1}{3} - \mu)^2(p_9 + \ldots + p_{20}) \\
&\quad + (\frac{2}{3} - \mu)^2(p_{21} + \ldots + p_{26}) + (1-\mu)^2(1 - \textstyle\sum_{m=1}^{26} p_m) \\
&= \sigma^2
\end{aligned}
$$

In practice, the p_i are estimated as sample fractions. For the n_h individuals with a common history follows

(B3) $E(\frac{1}{n_h}\sum_i \mu) = \mu_h$ and

(B4) $Var(\frac{1}{n_h}\sum_i (\frac{\sum 1}{3})) = \frac{1}{n_h}\sigma^2 = \sigma_h^2$

which yields the variance for both elements of the difference in (9). The variance of (9) then results from the sum of the two history-specific variances (B4) for treated and comparison units. Parallel to the derivation of the overall treatment effect (10) from the history-specific effect (9), the variance of (10) is a weighted sum (with squared weights) of the variance of (9).

Table 1. Composition of matched samples

		Training		Intervention Works	
		treated	**untreated**	**treated**	**untreated**
Sample A	N	121	6751	275	6757
	age	34.5	33.1	36.3	33.1
	%education[a]	91.7	80.7	64.0	80.7
	%female	56.2	53.0	40.4	53.0
	%married	66.9	65.8	67.6	65.6
Sample B	N	114	983	244	1354
Sample C	N	87	111	212	240
	[Individuals][b]		[110]		[211]

[a] Excluding individuals with only primary school attainment or less.
[b] Number of observations that the algorithm matched exactly once.

Table 2. Average post-treatment employment rate treatment effect by pre-treatment labor market history for Sample C – Treatment: Intervention Works

	treated units			comparison units				
job history	**N**	**rate[a]**	**std.err.**	**N**	**rate**	**std.err.**	**effect[b]**	**std.err.**
0000	5	0.333	0.189	6	0.400	0.219	-0.067	0.289
0002	1	0.000	0.000	1	0.667	0.471	-0.667	0.471
1111	16	0.813	0.098	19	0.729	0.111	0.084	0.148
1112	5	0.467	0.202	6	0.167	0.167	0.300	0.262
1122	6	0.222	0.150	6	0.333	0.192	-0.111	0.244
1222	4	0.500	0.250	4	0.833	0.186	-0.333	0.312
2000	1	1.000	0.000	1	0.000	0.000	1.000	0.000
2111	1	1.000	0.000	1	1.000	0.000	0.000	0.000
2211	4	0.167	0.144	4	0.667	0.236	-0.500	0.276
2221	1	0.000	0.000	1	0.333	0.471	-0.333	0.471
2222	168	0.183	0.027	191	0.333	0.036	-0.150	0.045
total[c]	212			240			-0.126	0.040

[a] Average employment rate in the three post-treatment quarters.
[b] Difference between rates of treated units and matched comparison units.
[c] Total effect is the weighted average of the effects for the individual histories using the treated units' sample fractions as weights.

Table 3. Average post-treatment employment rate treatment effect for subsamples of Sample B – Treatment: Intervention Works

Stratification by	Categories	treated units	matched comparison units	effect[a]	std.err.
Sample A	-	275	6757	-.285	.026
Sample B	-	244	1354	-.291	.031
Sample C:	-	212	240	-.126	.040
Gender	Men	123	133	-.236	.051
	Women	89	107	.026	.062
Date of Program Entry	≤ June 1994	116	137	-.135	.052
	≥ July 1994	96	103	-.115	.056
Program Entry & Gender	≤ June 1994 Men	66	73	-.295	.069
	≤ June 1994 Women	50	64	.076	.079
	≥ July 1994 Men	57	60	-.167	.073
	≥ July 1994 Women	39	43	-.038	.089
Labor market history	1111	16	19	.084	.148
	2222	168	191	-.150	.045
Labor market history & Gender	1111 Men	10	12	.117	.161
	1111 Women	6	7	.028	.274
	2222 Men	100	108	-.258	.057
	2222 Women	68	83	.010	.072

[a] Average employment rate in the three post-treatment quarters.

Table 4. Average post-treatment employment rate treatment effect for subsamples of Sample B – Treatment: Training

Stratification by	**Categories**	**treated units**	**matched comparison units**	**effect**[a]	**std.err.**
Sample A	-	121	6751	-.027	.046
Sample B	-	114	983	-.048	.049
Sample C:	-	87	111	.138	.059
Gender	Men	36	39	.148	.092
	Women	51	72	.130	.070
Date of Program Entry	≤ June 1994	38	52	.212	.088
	≥ July 1994	39	59	.080	.064
Program Entry & Gender	≤ June 1994 Men	15	17	.056	.156
	≤ June 1994 Women	23	35	.313	.104
	≥ July 1994 Men	21	22	.214	.094
	≥ July 1994 Women	28	37	-.020	.086
Labor market history	1111	24	34	.071	.115
	2222	32	43	-.077	.103
Labor market history & Gender	1111 Men	11	12	.045	.194
	1111 Women	13	22	.092	.129
	2222 Men	11	12	-.046	.192
	2222 Women	21	31	.093	.116

[a] Average employment rate in the three post-treatment quarters.

Table 5. Benefit receipt, sample C

	Intervention Works		
	treated	comparisons	difference
During 3 months BEFORE treatment			
men	52.03	35.34	+16.69
women	33.71	22.43	+11.28
During 3 months AFTER treatment			
men	60.16	11.28	+48.88
women	46.07	11.21	+34.86
During 3 months AFTER treatment			
men	55.28	6.77	+48.51
women	39.33	5.61	+33.72

Notes: % benefit recipients. The upper panel indicates benefit receipt (="yes") during at least two of the last three months preceding treatment. The middle panel indicates benefit receipt during at least two of the first three months succeeding treatment. The bottom panel indicates benefit receipt during at least two of the three months in each of the three quarters succeeding treatment.

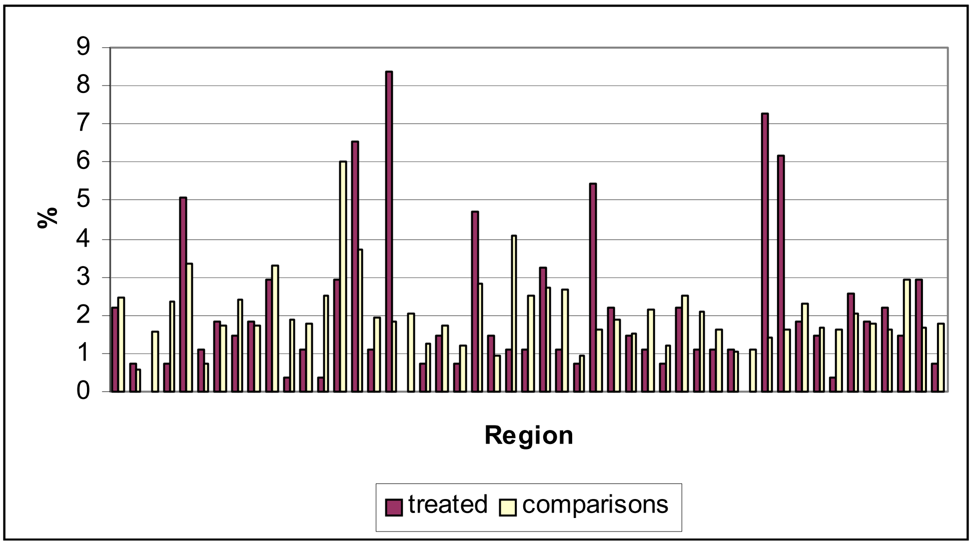
Table 6. Counterfactual treatment effects for samples C

Treatment	Weights	Effect[a]	Std.Err.	Interpretation
Intervention Works	Intervention Works	-.126	.040	Factual IW treatment effect
Intervention Works	Training	-.048	.064	Counterfactual IW treatment effect
Training	Training	.138	.059	Factual Training treatment effect
Training	Intervention Works	.089	.083	Counterfactual Training treatment effect
Intervention Works – Training	Intervention Works	-.218	.093	Differential treatment effect Intervention Works vs. Training
Training – Intervention Works	Training	.185	.087	Differential treatment effect Training vs. Intervention Works

[a] Average employment rate in the three post-treatment quarters.

<u>Figure 1. Distribution of region – Intervention works</u>

Sample A



Region = 49 voivodships.

Figure 2. Matching using a "moving window" in Sample B

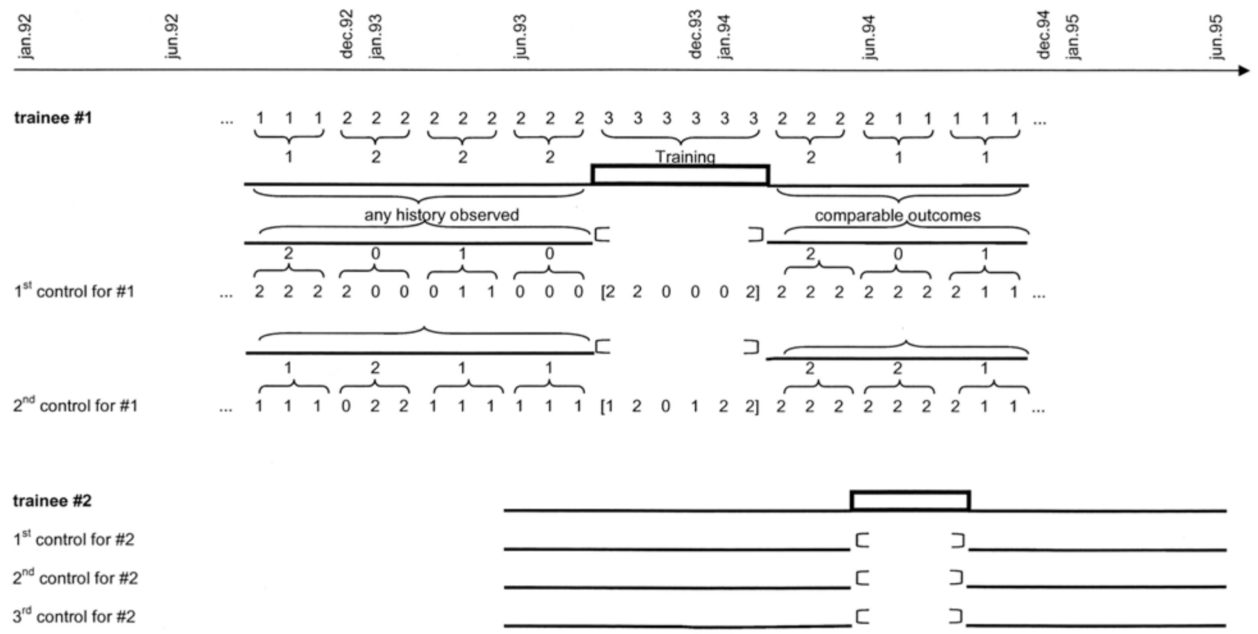


Figure 3. Matching over identical individual labor market histories using a "moving window" in Sample C

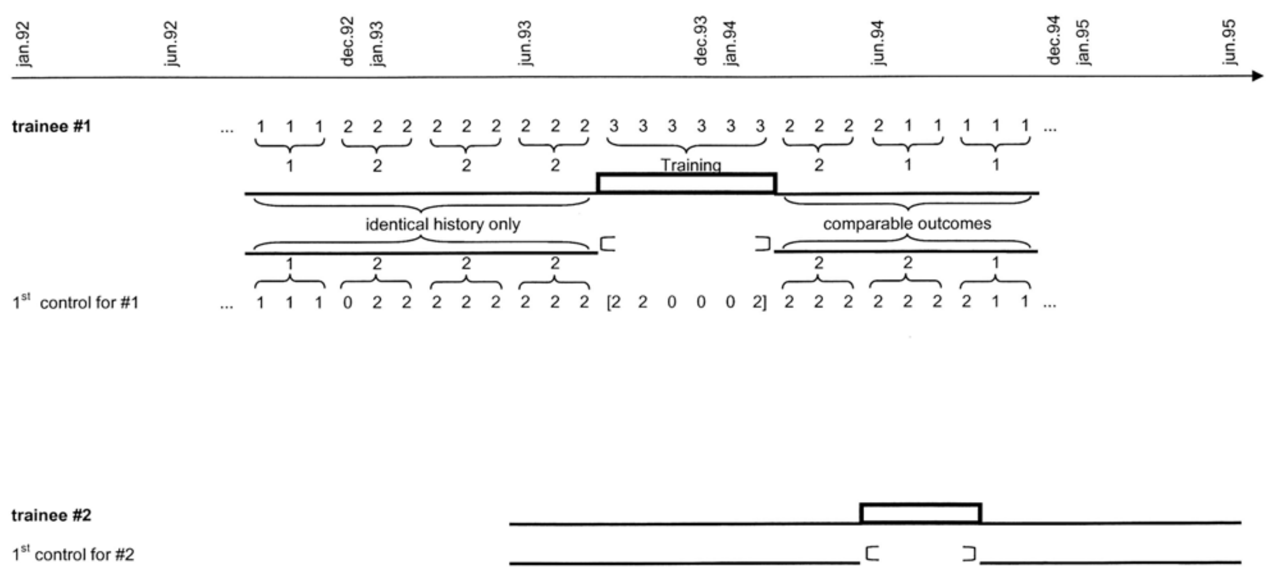
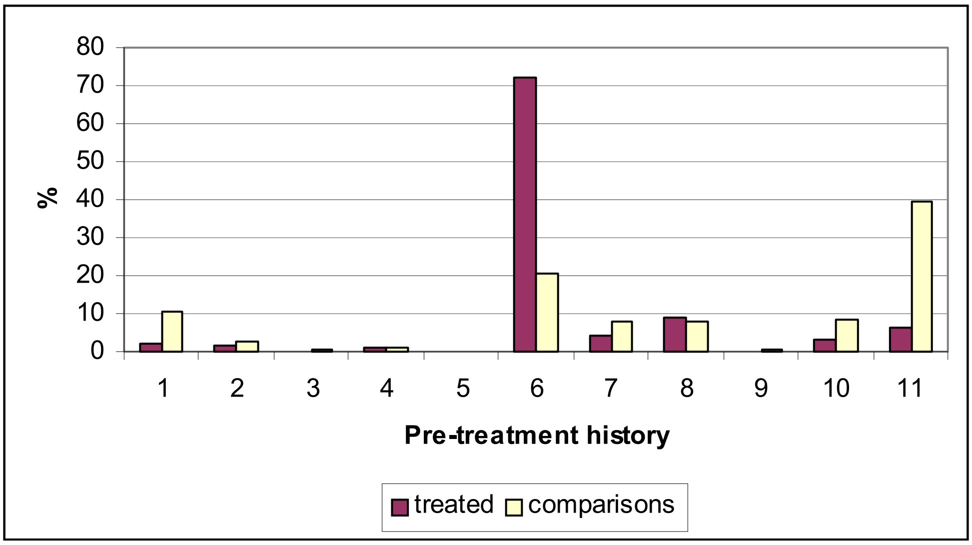
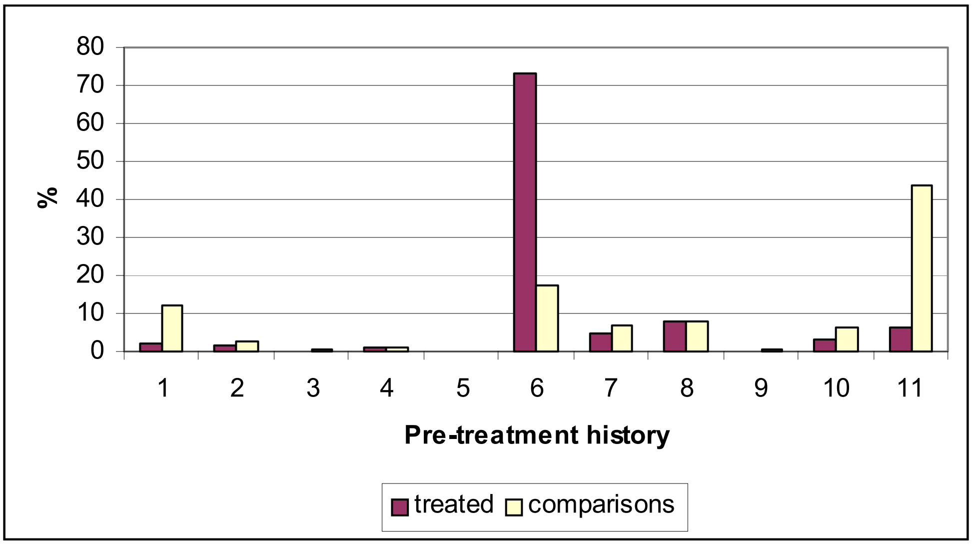


Figure 4. Distribution of pre-treatment labor market history by sample – Intervention Works

Sample A



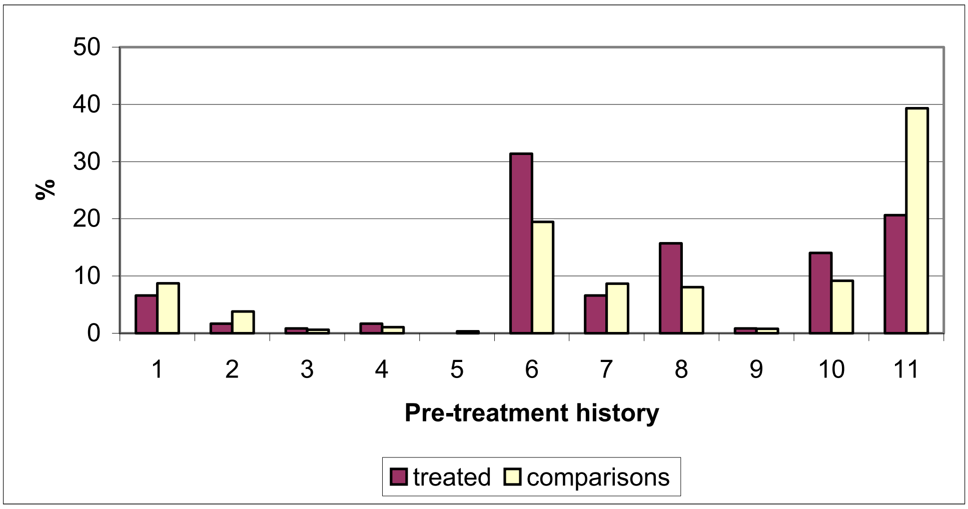
Sample B



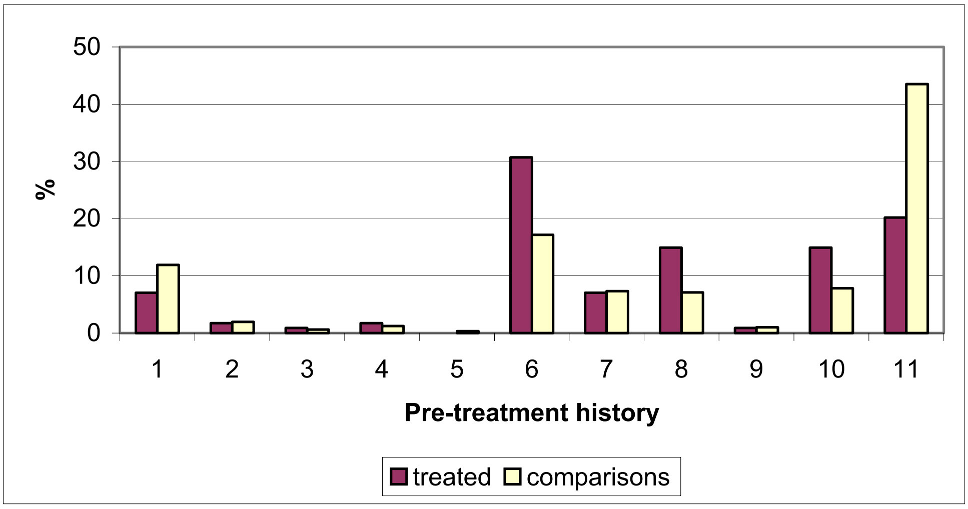
The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

Figure 5. Distribution of pre-treatment labor market history by sample – Training

Sample A



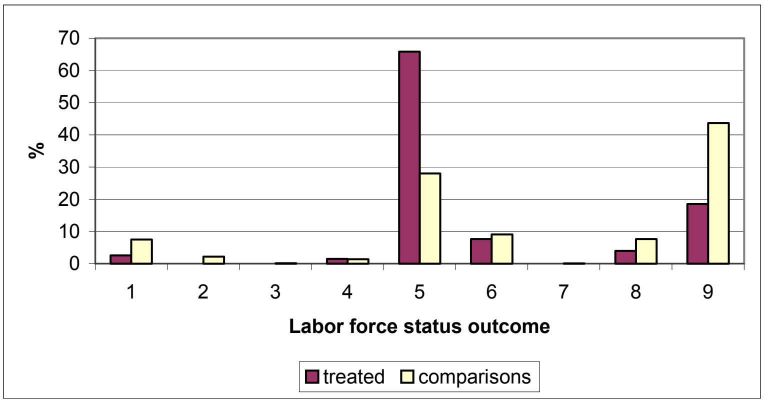
Sample B



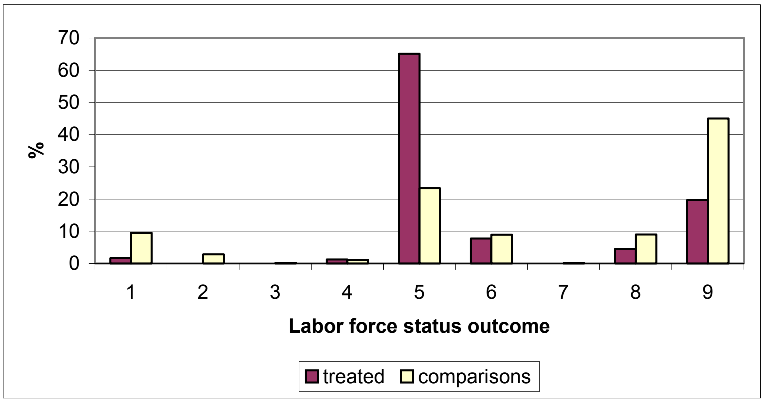
The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

Figure 6. Distribution of post-treatment labor market sequence by sample – Intervention Works

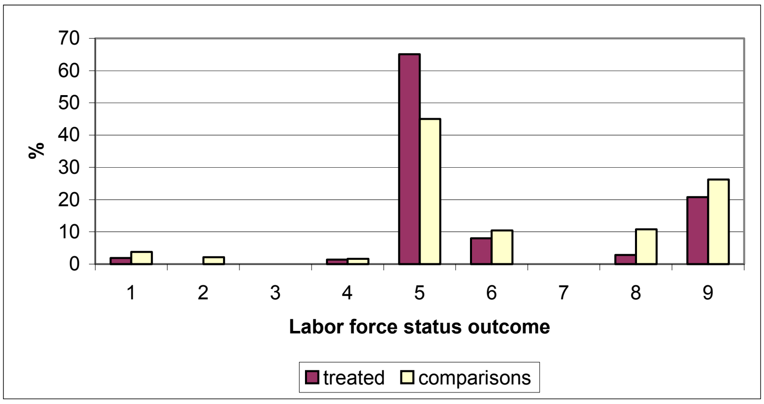
Sample A



Sample B



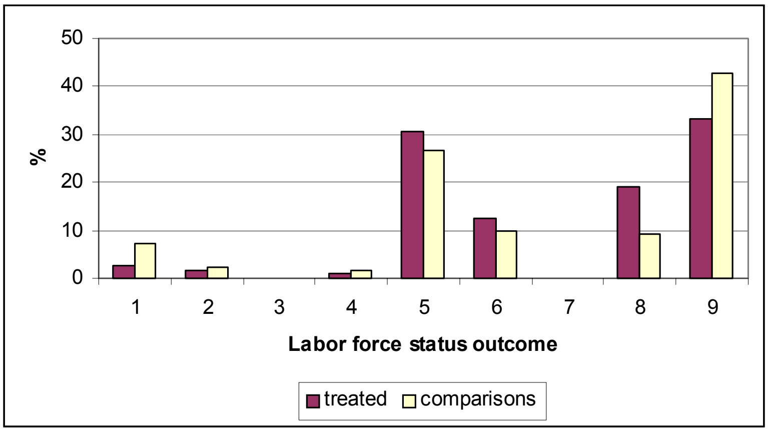
Sample C



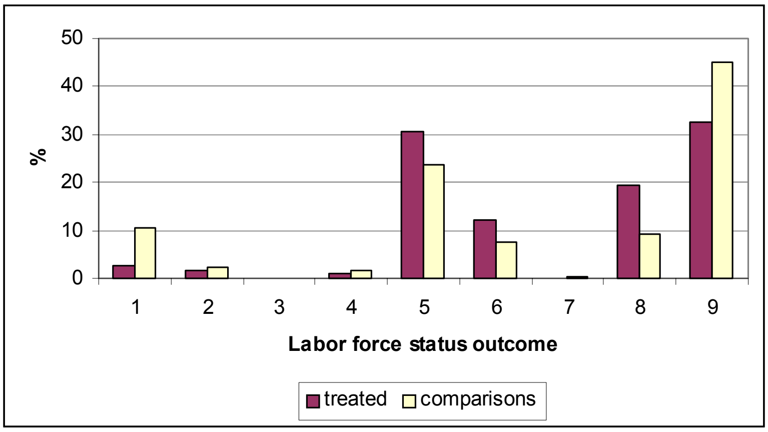
The 3^3 possible labor force status sequences are classified into 9 categories (see text and Appendix A).

Figure 7. Distribution of post-treatment labor market sequence by sample – Training

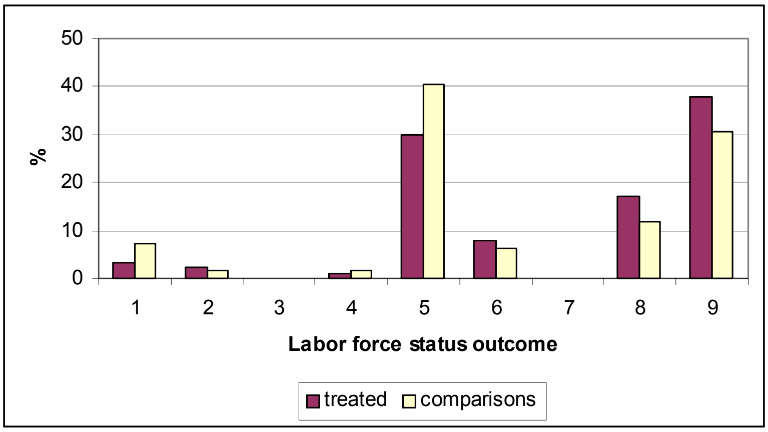
Sample A



Sample B



Sample C



The 3^3 possible labor force status sequences are classified into 9 categories (see text and Appendix A).