



LABORatorio R. Revelli
Centre for Employment Studies

Disentangling Treatment Effects of Polish Active Labor Market Policies: Evidence from Matched Samples

Jochen Kluge

AWI University of Heidelberg, and IZA Bonn

Hartmut Lehmann

Heriot-Watt-University Edinburgh, IZA Bonn,
WDI University of Michigan, Ann Arbor, and EERC Kiev

Christoph M. Schmidt

AWI University of Heidelberg, IZA Bonn, and CEPR London

Disentangling Treatment Effects of Polish Active Labor Market Policies: Evidence from Matched Samples

Jochen Kluge

AWI University of Heidelberg, and IZA Bonn

Hartmut Lehmann

*Heriot-Watt-University Edinburgh, IZA Bonn,
WDI University of Michigan, Ann Arbor, and EERC Kiev*

Christoph M. Schmidt

AWI University of Heidelberg, IZA Bonn, and CEPR London

This Version: January 2002

Abstract. This paper estimates causal effects of two Polish active labor market policies – Training and Intervention Works – on employment probabilities. Using data from the 18th wave of the Polish Labor Force Survey we discuss three stages of an appropriately designed matching procedure and demonstrate how the method succeeds in balancing relevant covariates. The validity of this approach is illustrated using the estimated propensity score as a summary measure of balance. We implement a conditional difference-in-differences estimator of treatment effects based on individual trinomial sequences of pre-treatment labor market status. Our findings suggest that Training raises employment probability, while Intervention Works seems to lead to a negative treatment effect for men. Furthermore, we find that appropriate subdivision of the matched sample for conditional treatment effect estimation can add considerable insight to the interpretation of results.

JEL Classification: C41, J68

Keywords: Active Labor Market Policy, transition, exact matching, propensity score

Corresponding Author:

Christoph M. Schmidt, Alfred Weber Institute, University of Heidelberg

Grabengasse 14, D-69117 Heidelberg, email: schmidt@uni-hd.de, Fax: +49-6221-543640

The authors are grateful to Boris Augurzky, and to Patrick Puhani and to the participants of the first joint IZA/WDI conference on "Labor Markets in Transition Countries" in May 2000 in Bonn, and to seminar participants at the University of Heidelberg for helpful comments. Research Assistance by Markus Müller is gratefully acknowledged. Lehmann and Schmidt are grateful to the Volkswagen Foundation for financial support within the project "Analysis of Labour Market Adjustment in Poland and Estonia with Large Micro Data Sets". Kluge acknowledges financial support from the *Cusanuswerk*.

1. Introduction

The evaluation of active labor market policy (ALMP) in the transition countries of Central and Eastern Europe faces serious methodical obstacles. Most importantly, studies typically have to rely on nonexperimental data, a feature they share with most evaluation studies on measures of active labor market policy in OECD countries. In fact, nonexperimental settings are still predominant in any European country study, as large-scale – or any – experimental studies similar to those conducted in the US have remained highly uncommon.

Apart from this more general drawback early evaluation studies on transition countries frequently had to be based on yet inadequate data: certainly, first of all, local national statistics offices had to gather experiences in generating data sets. Moreover, as the urge to evaluate programs already emerged almost simultaneously with the introduction of the data sets and the introduction of the policy measures themselves, early studies could not exhaust any long-term data. And yet another distinct feature of policy evaluation in a transition country is the need to control for the – in early years after transition – quickly changing macro environment, in particular if one aims at estimation of individual treatment effects.

The transition countries of Central Europe display a U-shaped pattern of output over the first years of transition, showing an initial contraction in economic activity after the onset of reform followed by, in the Polish case, robust expansion (cf. Blanchard 1997). The effectiveness of ALMP measures depends – *ceteris paribus* – on the tightness of the labor market and, therefore, on the point on the U-curve where the economy is located. Evaluating the effects of ALMP measures administered over several years without controlling for the large moves along the U-curve observed in Central European transition countries would severely bias the results.

This study focuses on the evaluation of active labor market policy in Poland, with an emphasis on two major points. First, with regard to the implicit missing data problem in any nonexperimental evaluation study, we explore the potential of different matching procedures to achieve covariate balance, and we demonstrate how in our case exact matching methods may in an intuitively appealing way resolve the dilemma of constructing an adequate counterfactual. To this end we discuss three stages of a matching procedure that is meticulously adapted to the specific nature of the data. Our

arguments are illustrated by comparing covariate balance and balance in estimated propensity scores – a summary measure of balance – across post-match samples.

Second, we discuss our evaluation results in detail, confirming earlier results on Polish ALMP (cf. Kluve, Lehmann and Schmidt 1999, Puhani 1998). We place particular emphasis on the necessity of considering subsets of the population of treatment units in the interpretation of results. We argue emphatically that a careful interpretation of results is as important as the devotion of effort to constructing an adequate comparison group, an idea that frequently seems to be overlooked in applied work. Specifically, we demonstrate that – even though an appropriate matching method does control for the relevant variables – once the comparison group is found, the analysis is not complete. Instead, pursuing the estimation of conditional treatment effects for appropriately defined subsamples may be useful to avoid otherwise misleading results.

The paper is organized as follows: Section 2 presents a brief description of the data and gives a short exposition of the evaluation problem, showing how matching on covariates and/or the propensity score can identify the treatment effect. Section 3 explains how our matched samples were constructed and to what extent the matching methods applied succeed in balancing observable covariates. Section 4 focuses on developing our matching estimator of treatment effects, on interpreting treatment effect estimates, and on the importance of conditioning treatment effect estimates on covariates for interpretation purposes. 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. From these questions, individual labor market histories in quarterly structure have been constructed. The individual histories cover the 56-month-period from January 1992 to August 1996. Yet, the retrospective data required considerable preparatory work.

First, out of an initial number of 48,385 observations 11,102 individual labor market histories lacked any entry, and were omitted from the analysis. The vast majority

of these are individuals who were inactive in August of 1996. From the remaining data we had to exclude both treatment participants with too early (before January 1993) or too late (after November 1995) treatment spells since in our econometric approach we condition on pre-treatment histories spanning one year and look at post-treatment labor market outcomes averaged over three quarters. Incomplete spells containing too little information were also excluded from the analysis.

Our analysis focuses on 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. Since we focus on two distinct ALMP programs, Training and Intervention Works, the resulting samples of treatment participants for both measures and their potential comparisons are substantially smaller than the initial data set. We discuss sample composition in more detail in section 3.1.

Secondly, in order to be able to handle such rich data, we had to condense the information contained in individual labor market histories. Monthly entries entail, e.g., "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 our estimation of individual treatment effects we consider two distinct measures of Polish ALMP: Training and Intervention Works¹. For more information on institutional details, on ALMP regulations and descriptions of courses we refer to earlier papers on the topic (Kluve et al. 1999, Puhani 1998, Góra and Schmidt 1998). For our purposes in this study it is mainly important to note the distinct nature of the two programs. Training is meant to enhance, or at least sustain, individual human capital during a period of unemployment. The Polish Training measure for the unemployed is

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 small sample sizes. Cf. also Kluve et al. (1999), Puhani (1998).

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 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 did, on the other hand, not entitle a person to renewed benefit payments since this training was done off-the-job.

2.2 Matching as a substitute for randomization

Program evaluation aims at estimating causal effects, i.e. changes in the variable of interest that are due to treatment participation. The formal setting is cast into the statistical "potential outcome framework" for causal inference based on Neyman (1923 [1990], 1935), Fisher (1935) and Rubin (1974, 1977). Let us consider a population indexed by i , and let Y_{i1} denote the variable of interest given individual i participated in a program, indicated by $D_i=1$. Likewise, let Y_{i0} denote the outcome if $D_i=0$, i.e. if individual i was not a participant, and define the single unit treatment effect as $\Delta_i=Y_{i1}-Y_{i0}$. However, outcomes Y_{i1} and Y_{i0} are "potential" in that we can never observe both of them simultaneously for one individual. The parameter of interest in nonexperimental studies is the mean effect of treatment on the treated population:

$$(1) \quad \Delta|_{D=1} = E(\Delta_i | D_i = 1) = E(Y_{i1} | D_i = 1) - E(Y_{i0} | D_i = 1)$$

The equation shows the inherent missing data problem, as we cannot observe the non-treatment outcome Y_{i0} for treatment participants $D_i=1$. We thus have to rely on

establishing a convincing substitute for $E(Y_{i0}|D_i=1)$ in equation (1) in order to identify the desired parameter.

In an experimental study randomization ensures that potential outcomes Y_{i1} and Y_{i0} are independent of treatment assignment D_i , i.e. $Y_{i1}, Y_{i0} \perp D_i$. Hence, program participants and comparison group do not systematically differ from each other, yielding the expectation of Y_{i0} for the comparison group as a substitute for the expectation of Y_{i0} of the treated group. Thus,

$$(2a) \quad E(Y_{i0} | D_i = 1) = E(Y_{i0} | D_i = 0) = E(Y_i | D_i = 0),$$

where Y_i is the actually observed value of the outcome variable, i.e. $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$. Thus, randomization ensures identification of the desired parameter $\Delta|_{D=1}$ from equation (1). Randomization also implies an assumption referred to as stable-unit-treatment-value assumption (SUTVA, see e.g. Rubin 1980): Potential outcomes for each individual are not related to the treatment status of other individuals, i.e. $Y_{i0}, Y_{i1} \perp D_j \forall i \neq j$.

Given a nonexperimental setting it appears appropriate to substitute for missing randomized-out controls by constructing a set of potential comparison units for whom we observe the same set of pre-treatment covariates X_i as for the treated units. The following proposition given in Rubin (1977) extends the above framework to nonexperimental studies:

If for each unit we observe a vector of covariates X_i , and $Y_{i0}, Y_{i1} \perp D_i | X_i$ holds $\forall i$, then the population treatment effect for the treated $\Delta|_{D=1}$ is identified: it is equal to the treatment effect conditional on covariates and assignment to treatment $\Delta|_{D=1, X}$ averaged over the distribution $X|D_i=1$.

Such a construction of counterfactual outcomes can only be sensible if conditioning is on variables which itself are not the outcome of treatment participation. Post-treatment employment success is a case in point: by matching those individuals who are or are not successful, the effect of treatment will necessarily be derived to be zero. Similar conceptual reservations would hold for characteristics of post-treatment jobs such as industry or working hours.

Consequently, conditional on observable covariates assignment to treatment can be considered as having been random, and unobservable characteristics possibly influencing treatment participation are ruled out. In fact, by this proposition comparing a program participant with a comparison individual displaying the same observable characteristics is like comparing the two in a randomized experiment. We thus merely need to estimate $E(Y_{i0} / X_i, D_i=0)$, so that

$$(2b) \quad E(Y_{i0} | D_i = 1) = E_X (E(Y_{i0} | X_i, D_i = 0) | D_i = 1),$$

identifying the mean effect of treatment on the treated of equation (1) for a nonexperimental setting: constructing the appropriate weighted average over conditional (on X) no-treatment outcomes mimics randomization by balancing all relevant covariates.

Ideally, in order to implement a procedure for estimating the conditional treatment effect $\Delta_{D=1,X}$, we could simply match treated and comparison units on their covariate vector X_i . While exact matching on X_i achieves an exact balancing of attributes, it suffers from the fact that X_i might be of high dimension or contain continuously-distributed variables, so that some treated units might not find comparisons. To avoid the problem of matching on a high-dimensional X_i , the method of propensity score matching has been proposed by Rosenbaum and Rubin (1983). Define the propensity score as $p(X_i) = Pr(D_i=1|X_i) = E(D_i|X_i)$, i.e. the conditional probability of receiving treatment given a set of covariates. Then the conditional independence result from above extends to the propensity score: $Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i | X_i \Rightarrow Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i | p(X_i)$.

The reduced dimension comes at a cost, however. The propensity score is not known and has to be estimated. Also, in samples of limited size, for some i and j it may occur that $p(X_i) = p(X_j)$ even if $X_i \neq X_j$, resulting in imperfect balancing of the distributions of covariates. Thus, the small sample performance of propensity-score matching might be quite dismal. In fact, the literature indicates that the trade-off between exact matching and propensity score matching² is one of truly empirical nature: The decision for one approach or the other should depend heavily on the data, e.g. the number of

² In practice, matching algorithms are manifold, including e.g. exact matching, matching within calipers (fixed or flexible), minimum-distance matching, or optimal full matching minimizing total distance. For further reference see Gu and Rosenbaum (1993), Rosenbaum (1995), and Augurzky (2000).

observations, the dimension of X , time structure of variables, etc., and certainly it should depend on what the researcher believes (and justifies) to be the adequate *modus operandi* in each specific case.

Angrist and Hahn (1999) make this point forcefully by stating that existing theory provides little in the way of specific guidelines as how to choose between the two. On the one hand Hahn (1998) proves that exact matching is asymptotically efficient while propensity score matching is not, and concludes that asymptotic arguments would appear to offer no justification for anything other than full control for covariates. On the other hand Angrist and Hahn (1999) show that in some plausible scenarios estimators controlling for the propensity score can be more efficient than exact-matching estimators. The latter seems to be valid in particular when cell-sizes are small, the explanatory value of the covariates is low conditional on the propensity score, and/or the probability of treatment is far from $\frac{1}{2}$.

Still, what counts in practice is how well balance is achieved, so that the researcher can indeed "compare the comparable" (Heckman et al. 1997). Any matching procedure allowing for any distance in either X or $p(X)$ must be aware of that. And including a weak predictor of $p(X)$ into the estimation might be more harmful than covariate matching on a reduced set of comparisons. Thus, both Dehejia and Wahba (1998) – based on an empirical study – and Augurzky and Schmidt (2000) – based on a simulation study – argue that it is more important to achieve balance of relevant covariates rather than painstakingly modeling the selection process.

3. Analyzing Matched Samples

3.1 Composition of Matched Samples

For each of the two measures under scrutiny – Training and Intervention Works – we start the construction of matched samples from an initial sample consisting of treated individuals and untreated potential comparison individuals, where every observation is required to have at least one spell of unemployment. From this starting point we subsequently impose stronger restrictions on X (i.e. enlarge the dimension of the matching criteria) step-by-step, in order to obtain three samples of matched treatment-comparison units for each of the two measures:

Sample A: A comparison unit is matched to a treated unit if his or her labor market history is observed without substantial gaps from a year before up to the beginning of treatment and from the end of treatment until 9 months later. None of the observed individual characteristics is used as a matching criterion.

Sample B: A comparison unit is matched to a treated unit if requirement (A) is met, and if he or she is identical in observable characteristics age, gender, education, marital status, and region.

Sample C: A comparison unit is matched to a treated unit if requirements (A) and (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

Samples (A) through (C) are constructed applying an exact-matching-within-calipers algorithm. For all three samples, if a treated individual finds any matching partner among the potential comparisons, this observation is retained. All algorithms allow for an oversampling procedure, i.e. a treated unit may be assigned more than one comparison unit. While we could have sharpened the matching criteria in a different order, this sequence 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 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 resulting sample sizes, as well as means of relevant variables. We observe that there is a reduction in the number of treated units who find matching

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 $3528 \cdot 3^4 = 285,768$ cells for sample (C).

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.

< Table 1 about here >

Throughout, we focus our attention on exact matching procedures. In sample (B), the number of matching variables is limited, and they are all categorical variables. Moreover, exact matching performs quite well: despite the substantial number of cells, approximately 9 out of 10 of treated units find a comparison unit. With regard to sample (C), our exact matching approach is a very practical device to account for the pre-treatment employment sequence. Further illustration is provided in the next sections.

3.2 Timing of interventions

In our preferred sample (C) we require treated and matched comparison units to display an identical pre-treatment history. To achieve comparability across 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 exactly is not used in matching. Moreover, to allow an assessment of post-treatment labor market performance, we require for treated units and all comparison samples that we observe a post-treatment sequence of labor force status variables in the nine months after treatment. In accordance with our preparatory data work, we condense the monthly information for treatment units to a sequence of three quarters of a multinomial outcome variable (0,1,2) denoting labor force status (out-of-the-labor force, employed, unemployed).

Correspondingly, for those comparison units eventually matched to the treated units, a comparable three-quarter post-treatment multinomial sequence of labor force status is computed as well, again starting at the exact point in time when the treatment spell of the corresponding treated unit ended. Our analysis thus incorporates individual treatment duration by conditioning on a complete (i.e. without major gaps) pre-

treatment labor market history being observed before month "start" and comparing labor force status outcomes after month "stop". Thus, treated units and matched comparison units are always being compared during the same period. Figure 1a illustrates this procedure for samples A and B, in which the timing structure is considered, but the contents of individually matched labor force status histories does not matter. Figure 1b proceeds to depict the case for inclusion of exact pre-treatment histories in matching for sample (C)

< Figures 1a,1b about here >

We thus take advantage of the specific nature of the data with monthly information on employment status for a 56-month period, considering the exact timing of "start" and "stop" of treatment – a feature that is neither common nor possible in many studies, even those focussing on duration data. Moreover, given the rapid upward moves of the Polish economy along the positive section of its U-shaped curve of output between 1992 and 1996, we can assume that labor market tightness has increased in Poland in the reported period. Hence, the fact that we are able to compare treated and comparison units individually at the same point of time seems particularly valuable.

There might be other ways to solve the crucial problem of finding the "starting point of treatment" for comparison units. In principle, one could first match on characteristics X or the propensity score conditioned on characteristics, $p(X)$, and then directly impose requirements on comparable timing. A procedure following such a "partial balancing score" is for instance used by Lechner (2000). It seems more natural to us, however, to incorporate timing as a principal component of matching.

3.3 Covariate balance

In section 2.2 we have emphasized that balance in all relevant factors – observed as well as unobserved – is the principal objective in experiments, and in its observational counterpart, the matching approach. In this section we examine whether the particular matching procedures we applied here indeed succeed in balancing the distributions of pre-treatment covariates between treatment units and their comparisons. Figures 2 and 3 show the distributions of the two principal covariates age and region for treated and comparison units when matching is according to requirements (A) and the analyzed treatment is Intervention Works. By contrast to sample (A), samples (B) and (C) match

on these individual characteristics. The figures illustrate by how much matching on the correct timing alone would miss out on balancing individual characteristics.

< Figure 2 about here >

Figure 2 shows that if not accounting for age, the young would be over-represented among the comparisons, and the mature (35-50, say) workers would be over-represented among the treated units.

< Figure 3 about here >

Figure 3 plots the frequency distribution for the 49 Polish voivodships. Including regional indicators among the matching covariates is firmly advocated by Heckman et al. (1997) in order to control for the local labor market. This is the more imperative in the Polish case, since local labor market conditions are quite heterogeneous in any typical transition country. The matching criteria for samples (B) and (C) achieve complete balance – besides oversampling of comparison units – in the distribution of voivodships for treated and comparison units, while sample (A) displays considerable imbalance. Thus, if regional information were left out of the matching algorithm, regional balance would not be assured.

With respect to further socio-demographic characteristics, 59.6% of Intervention Works participants are male, while there are only 47% men in comparison sample (A). Regarding the three education categories, the middle category comprises 63.6% of Intervention Works participants, and there is only one single individual out of the 275 treated (=0.36%) in the top category. Among comparison units in (A), 2.4% and 78.4% are in the top and middle categories, respectively. Table 1 shows that sample (B) and in particular sample (C) achieve balance in terms of sex and education.

3.4 Pre-Treatment Histories

The literature on program participation has always been concerned with the focal problem of controlling for observable characteristics, unobserved heterogeneity, and selection bias. Mainly affecting a difference-in-differences estimation approach, Ashenfelter (1978) pointed to a potentially serious limitation of this procedure when he observed a relative decline in pretreatment 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 on earnings changes as the determinants of participation. This line of thought was a natural consequence of Ashenfelter's discovery and resulted in analyses using earnings histories to eliminate differences between participants and nonparticipants⁴. 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. 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.

We consider the 12-month labor market history of every single treated unit directly preceding the exact point in time – i.e. month – that the individual entered the program. As for the post-treatment outcomes, we condense the monthly information to a sequence of four quarters of a multinomial outcome variable (0,1,2) denoting labor force status (out-of-the-labor-force, employed, unemployed). For each treated unit in succession, the matching algorithm for sample (C) computes labor market histories for all potential comparison units at this point in time and matches those units who – in

⁴ Heckman and Smith (1999) attribute this emphasis also to the limited data available to "early analysts".

addition to the correspondence in the other covariates – display identical "pre-treatment" histories. For illustration see Figure 1b.

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). 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 balanced distributions – and only for that purpose – we classify these 81 sequences into 11 categories (see Appendix A), so that on the abscissa the low categories contain "inactive" sequences (mostly '0's), the middle categories comprise "unemployed" sequences ('2's), and the high categories represent "employed" sequences ('1's). Categories 1, 6, and 11 exclusively embody the straight sequences (i.e. "0000", "2222", and "1111", respectively).

< Figures 4 and 5 about here >

Thus, of the three peaks we observe in most of the graphs in figures 4 and 5, the left peak represents the area of "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 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.

3.5 Propensity score balance

The preceding sections were concerned with balance in selected individual characteristics. It is instructive to also provide a *summary measure of balance*, the propensity score. While the estimation of propensity scores is usually a principal step in

the construction of matched samples – with the hope that the resulting matched sample displays a balance in all relevant characteristics but no possibility to test this presumption – we can use our samples to directly analyze balance in the propensity score. Correspondingly, we predict post-match propensity scores for samples (A) and (B), based on estimates derived from sample (A). We follow a probit specification with interaction terms between some of the covariates,

$$(3) \quad P(D = 1 | X) = \Phi(\alpha_0 + \alpha_1 X + \alpha_2 X \otimes X)$$

where Φ denotes the cumulative normal density function, X is the vector of covariates, and $X \otimes X$ indicates all relevant interactions across covariates. Regressors comprise indicator variables capturing age, education, gender, and region. Moreover, corresponding to the condensation of pre-treatment labor market histories into 11 distinct "types" in section 3.4, there are 10 indicators of pre-treatment history among the regressors. Finally we interacted age, gender and education in a saturated fashion.

This model is estimated using the treatment units (yielding the value "1") and comparison sample (A) providing the "0" observations. Note that we observe both the individual characteristics and the pre-treatment histories also with comparison sample (A), although this information is utilized only in the construction of comparison samples (B) and (C), respectively. The resulting coefficients are employed to predict propensity scores in samples (A) and (B). Figures 6 and 7 document the distribution of propensity scores in these comparison samples – relative to the corresponding distribution among treatment units – for the two measures under study.

< Figures 6 and 7 about here >

Note that the density for treated units is not scaled relative to the number of observations in the comparison pool, so that the figure depicts the distribution of scores rather than the proportion of treated units to comparison units. In both figures 6 and 7 the comparison units gather at the low end of the estimated score. Whereas for Intervention Works treated units are distributed rather evenly, with the peak to the low end and then slightly declining towards the upper tail, the majority of treated units for Training also displays relatively low scores, with an overall distribution quite close to

that of comparison units. We find relatively little change in balance from (A) to (B) for both Training and Intervention Works. For Training the distributions are rather balanced – for Intervention Works, however, the substantial imbalance in pre-treatment histories clearly finds expression in the score distributions for (A) and (B) that do not yet control for this imbalance.

4. Empirical results

4.1 Distributions of outcomes

To illustrate the substantial heterogeneity of labor market outcomes following Intervention Works and Training, Figures 8 and 9 plot distributions for the post-treatment employment success for treatment units and the comparisons 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 1a, 1b). 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, low categories contain "inactive" sequences (category 1="000"), middle categories include "unemployed" sequences (category 5="222"), and high 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.

< Figure 8 about here >

Looking at the Intervention Works samples in Figure 8, 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). For our preferred comparison 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. Attributing the most reliable results to sample (C), we would conclude that during the 9 months directly succeeding participation in Intervention Works the treated units on average were

marginally – possibly insignificantly – less successful in finding employment than the comparison units.

< Figure 9 about here >

For the Training samples shown in Figure 9 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. We would thus attribute a slightly – possibly insignificant – positive treatment effect to Training.

Taken together, Figures 8 and 9 display three 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, 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 result in a far too negative estimate of treatment effects. Third, across comparison units and treated units we observe clearly more "successful" outcomes for Training than for Intervention Works. This is not surprising, as we noticed a similar relation for pre-treatment labor market history distributions (Fig. 4 and 5).

In Figures 10 and 11 we address the idea that participation in Intervention Works might primarily be a vehicle to renew eligibility for unemployment benefits. Recall that according to Polish ALMP regulations Intervention Works renews benefit receipt eligibility, whereas Training does not. Figures 10 and 11 perform a simple before-after comparison of the variable "unemployment benefit receipt" for both ALMP measures, and for men and women separately. The top panel of each figure 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 plots 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 focus on sample (C) for both measures.

< Figure 10 about here >

Figure 10 shows for Intervention Works 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.

< Figure 11 about here >

The situation for the Training sample is quite different. As Figure 11 shows, unemployment benefits do play some role for both treated and comparison units during the one quarter directly before and after participation, at least for the males. However, in the medium run this effect diminishes, and only very few observations in the treatment and comparison group display benefit receipt for the whole 9-month period following treatment. This pattern is even less pronounced for women than for men.

As a result, figures 8 through 11 indicate that individuals involved in Training measures seem to be generally more successful before and after the treatment than those participating in Intervention Works. However, these patterns are difficult to reconcile on the basis of a more favorable impact of Training. Rather, this simple evidence suggests that substantial benefit churning seems to take place in the case of Intervention Works, but not in the case of Training.

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 $E(Y_{qi}^0 | D_i=1)$. 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 $\mathbf{1}(\cdot)$, these employment rate outcomes are $\frac{1}{3} \sum_q \mathbf{1}(Y_{qi}=1)$.⁵ Observed outcomes for individual i can then be

written as

$$(4) \quad \frac{1}{3} \sum_q \mathbf{1}(Y_{qi}=1) = \frac{1}{3} (D_i \sum_q \mathbf{1}(Y_{qi}^1=1) + (1-D_i) \sum_q \mathbf{1}(Y_{qi}^0=1)) \quad ,$$

⁵ 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.

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

$$(5) \quad \Delta_i = \frac{1}{3} (\sum_q \mathbf{1}(Y_{qi}^1 = 1) - \sum_q \mathbf{1}(Y_{qi}^0 = 1))$$

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

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

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 $D=1$,

$$(7) \quad M = \sum_s w_s E(\Delta | s, D = 1),$$

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,

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

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

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

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,

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

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

In this section we analyze the treatment effect estimates which we obtain by applying the estimator developed in the previous section. Table 2 presents average treatment effects on the post-intervention employment rate for Intervention Works sample (C). The structure of the table shows how the total treatment effect (-.126) is being 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 unit. The total effect is the weighted average of the history-specific effects using the treated units' sample fractions as weights.

< Table 2 about here >

Besides treatment effect calculation Table 2 shows 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 which we already noticed in the figure. The total treatment effect casts a rather negative picture on the Intervention Works program, suggesting that participation tends to lower post-treatment employment prospects. In principle, this finding would conclude our analysis: we have described the nonexperimental context of the study, we have shown by what means we overcome the problem of constructing the desired counterfactual, and we have applied the appropriate estimation methods in order to obtain credible treatment effect estimates. As far as the data permit, the causal effect of Intervention Works participation is identified. Or is it?

In fact, looking at Table 3 we find that there may be more to it. First, we report treatment effect estimates for comparison samples (A) and (B) obtained by taking sample averages over the average employment rate in the three post-treatment quarters. The estimates are far more negative than the one obtained using sample (C), clearly reflecting the over-representation of "successful" labor force status sequences in the respective comparison samples (cf. Fig. 4 and 8). Furthermore, in accordance with our discussion of expression (7), in Table 3 we subdivide the matched Intervention Works comparison sample (C) with respect to various covariates, and we compare the conditional treatment effect for the subsample to the full sample estimate. Even a simple subdivision 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.

< Table 3 about here >

The next step is to further refine cells and to classify 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 the time period distinction leads to the opposite result, but both the positive effect of the first half and the negative effect of the second half are small and insignificant. This also points to the fact that, as we increase the number of subdivisions, subsample sizes decrease and standard errors increase.

Classification by labor market history allows us to look at the two major labor force status sequences that drive the peaks from Figures 4 and 5. 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 classify 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 subdivisions for Training. Both treatment effect estimates from comparison samples (A) and (B) suggest an insignificantly negative effect of Training participation, while the estimate obtained from sample (C) indicates that Training raises the individual employment probability by 13.8%. This sudden switch of signs is in line with our observations drawn from Figure 9. Further looking at comparison sample (C), we conclude that in the case of Training a classification by gender does not seem to add any insights to the interpretation: Treatment effects for men and women are almost identical. While a categorization by gender and date of program entry shows contradictory results (upward for men, downward for women from one period to the other), the number of

observations per subsample is in fact too small to draw any firm conclusions. Looking at a classification by labor market history, once more we find the "peaks" from Figure 5, indicating here that the share of "1111" sequences is almost as large as the that of "2222" sequences. Again, subsample sizes are quite small for interpretation purposes.

< Table 4 about here >

From these calculations results the observation that an appropriate subdivision 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. Thus, while the recognition of the principal idea that treatment effects are heterogeneous across the population has led to the development of sophisticated econometric methods for constructing convincing counterfactuals, it is easy to forget the necessity to stratify the sample appropriately in order to interpret the results in economically meaningful terms. Thus, controlling for observable characteristics in establishing the statistical model does not seem to be sufficient – it appears to be good advice to re-consider the same observable characteristics (which we already controlled for) when analyzing the empirical results. This recommendation seems imperative if one wants to assess for example targeting issues: bad targeting of programs is often claimed to be one reason for disappointing treatment effects. In our particular application, Intervention Works has been uncovered as an extremely disappointing measure in the case of men – a result that would have remained hidden, had we not pursued an appropriate sample split.

Of course, these negative treatment effects could be explained by other factors than poor targeting. Stigma is often given as a reason why participants of an employment program like Intervention Works perform worse in the labor market than non-participants.⁶ 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

⁶ A large part of the intervention works jobs are actually in the public domain, i.e. we can also think of this scheme as a public employment program.

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 the least 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.

However, more work is needed to determine firmly the factor(s) that drive the poor labor market performance of males after their participation in Intervention Works comes to an end. For example, the fate of female participants after the end of the subsidized job needs to be more thoroughly analyzed. Specifically, one needs to ask whether female participants are more likely to be kept on by employers or whether they find regular jobs elsewhere more readily than men because their characteristics are better than those of men, i.e. because the targeting criteria are different for men and women. It could also be that women who participate in Intervention Works are selected into jobs that are more conducive to prolonged job matches because demand in these jobs is strong (e.g. nursing jobs).

In addition to displaying the treatment effects by sample and subdivision, Table 5 presents treatment effect estimates for comparison 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 .

< Table 5 about here >

The estimate in the second line of Table 5 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 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

In this paper we have analyzed treatment effects of two Polish measures of active labor market policy: Training and Intervention Works. The analysis was based on matched samples to overcome the inherent evaluation problem of constructing a credible counterfactual in a nonexperimental setting. We have seen how matching methods can solve this problem by balancing distributions of relevant covariates. Matching methods can be based on exact-covariate-matching, propensity score matching, or a combination of both (partial score). We have argued that on both theoretical and above all empirical grounds the decision for one approach or the other depends heavily on the data.

We have illustrated our own approach to the data by the construction of three different comparison samples using exact-matching-within-calipers, imposing increasingly stricter preconditions. Figures 1 to 5 have depicted how strong requirements, i.e. a more detailed match on observable characteristics substantially improve the balancing of covariates, and thus the quality of the match. As long as sample sizes do not decrease considerably, such a procedure appears promising. We have illustrated the balancing property of our exact matching approach using the estimated propensity score as a summary measure of balance.

The estimation of the treatment effect is based on a history-specific generalized difference-in-differences estimator. Our estimates suggest that, while Training seems to clearly enhance individual employment prospects, Intervention Works participants fare substantially worse than their comparisons. This is in line with previous findings (cf. Kluve, Lehmann and Schmidt 1999, Puhani 1998). However, we do point to the fact that appropriate subdivision of the matched sample can add considerable insight to the interpretation of results. In our study, for instance, we find that the overall negative treatment effect of Intervention Works is almost exclusively due to the dismal employment performance of male participants, while women do neither gain nor lose anything by participating. From an empirical point of view, we thus doubt that controlling for covariates in constructing the counterfactual is sufficient to account for the heterogeneity of treatment effects – appropriate subdivision of the matched sample may often add clarity to the economic interpretation.

References

- Angrist, Joshua D., and Jinyong Hahn (1999), "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *NBER Technical Working Paper 241*, Cambridge, MA.
- Ashenfelter, Orley (1978) "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics* 60, 47-57.
- Ashenfelter, Orley, and David Card (1985) "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics* 67, 648-660.
- Augurzky, Boris (2000), "Optimal Full Matching", *Dept. of Economics Discussion Paper 310*, University of Heidelberg.
- Augurzky, Boris, and Christoph M. Schmidt (2000), "The Propensity Score: A Means to an End", *Dept. of Economics Discussion Paper 334*, University of Heidelberg.
- Bassi, Lauri J. (1983) "The Effect of CETA on the Post-Program Earnings of Participants", *The Journal of Human Resources* 18, 539-556.
- Blanchard, Olivier (1997), *The Economics of Post-Communist Transition*, Oxford: Clarendon Press.
- Card, David, and Daniel Sullivan (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica* 56, 497-530.
- Dehejia, Rajeev H., and Sadek Wahba (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies", *NBER Working Paper 6829*.
- Fisher, Ronald A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.
- Góra, Marek and Christoph M. Schmidt (1998), "Long-term unemployment, unemployment benefits and social assistance: The Polish experience", *Empirical Economics* 23, 55-85.
- Gu, Xing Sam, and Paul R. Rosenbaum (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms", *Journal of Computational and Graphical Statistics* 2, 405-420.
- Hahn, Jinyong (1998), "On the Role of the Propensity Score in the Efficient Semi-parametric Estimation of Average Treatment Effects", *Econometrica* 66, 315-332.
- Heckman, James J., Hidehiko Ishimura and Petra E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies* 64, 605-654.

- Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in: Ashenfelter, Orley and David Card (eds.): *Handbook of Labor Economics*, vol. III, Amsterdam et al.: North-Holland.
- Heckman, James J., and Jeffrey 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.
- Kluve, Jochen, Hartmut Lehmann, and Christoph M. Schmidt (1999), "Active Labor Market Policies in Poland: Human Capital Enhancement, Stigmatization, or Benefit Churning?", *Journal of Comparative Economics* 27, 61-89.
- LaLonde, Robert J. (1986), "Evaluating the econometric evaluations of training programs with experimental data", *American Economic Review* 76, 604-620.
- Lechner, Michael (2000), "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources* 35, 347-375.
- Neyman, Jerzy (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, Jerzy (1935), with co-operation by K. Iwazskiewicz, and S. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation", (with discussion), *Supplement to the Journal of the Royal Statistical Society* 2, 107-180.
- Puhani, Patrick A. (1998), "Advantage through Training? A Microeconomic Evaluation of the Employment Effects of Active Labour Market Programmes in Poland", *ZEW Disc. Paper* 98-25, Mannheim.
- Rosenbaum, Paul R. (1995), "*Observational Studies*", New York: Springer Series in Statistics.
- Rosenbaum, Paul R., and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 41-55.
- Roy, Andrew D. (1951), "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers* 3, 135-146.
- Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66, 688-701.
- Rubin, Donald B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics* 2, 1-26.

Rubin, Donald B. (1980), Comment on “Randomization Analysis of Experimental Data: The Fisher Randomization Test” by D. Basu, *Journal of the American Statistical Association* 75, 591-593.

Schmidt, Christoph M. (1999), “Knowing What Works – The Case for Rigorous Program Evaluation”, *IZA Discussion Paper 77*, Bonn.

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)

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

and

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

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

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

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

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
Initial Sample	Observations	121	7177	275	7177
Sample A	Observations	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	Observations	114	983	244	1354
	age	34.0	33.0	36.0	34.7
	%education	93.9	98.9	69.3	87.4
	%female	56.1	62.1	40.6	51.9
	%married	65.8	23.2	70.5	77.8
Sample C	Observations	87	111	212	240
	[Individuals] ^b		[110]		[211]
	age	33.4	33.8	36.0	35.2
	%education	96.6	97.3	71.2	74.2
	%female	58.6	64.8	42.0	44.6
	%married	67.8	70.3	70.3	70.4

^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 comparison sample C – Intervention Works

job history	treated units			comparison units			effect ^b	std.err.
	N	rate ^a	std.err.	N	rate	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 – Intervention Works

Subdivision 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 – Training

Subdivision 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. 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.

Figure 1a. Matching applying a "moving window" in samples (A) and (B)

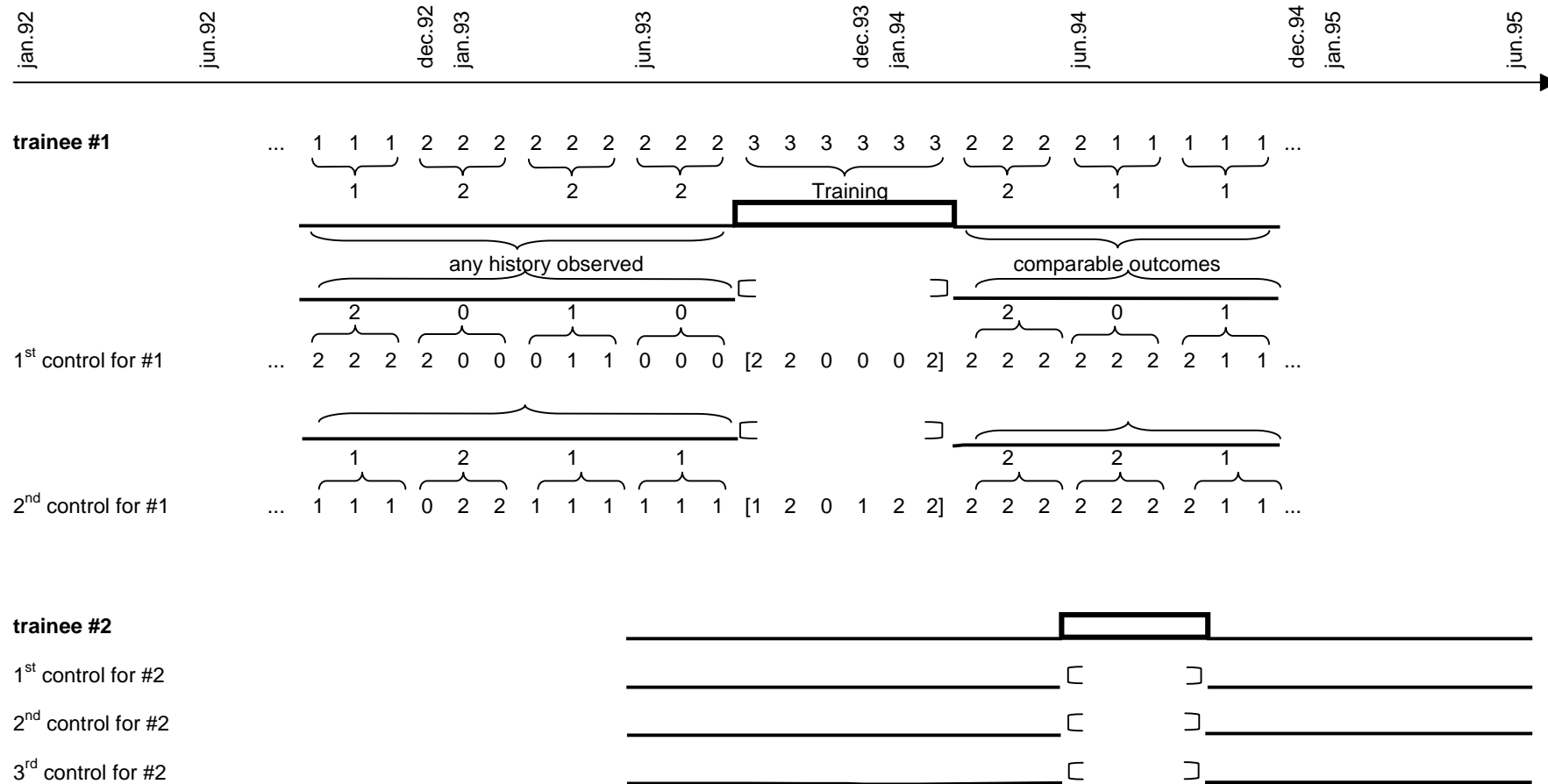


Figure 1b Matching over identical individual labor market histories applying a "moving window" in sample (C)

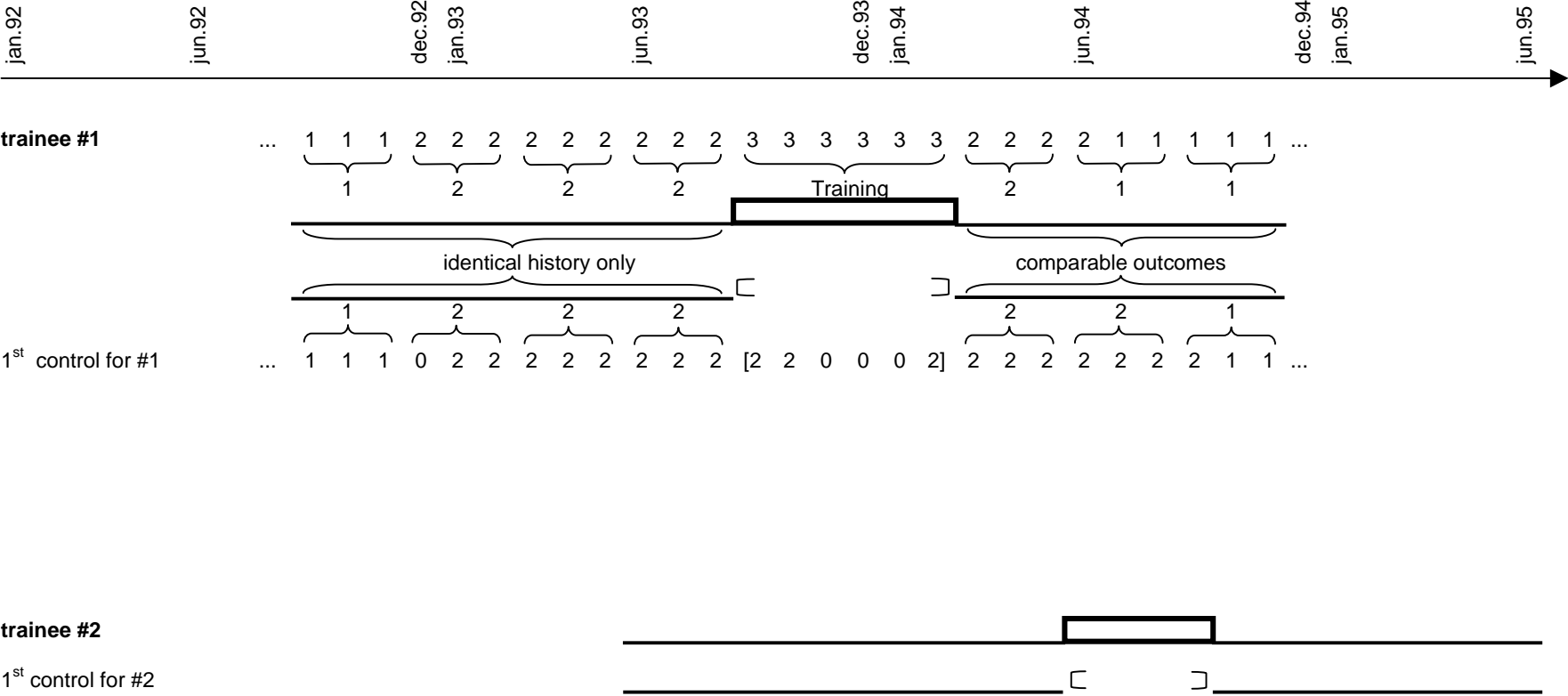
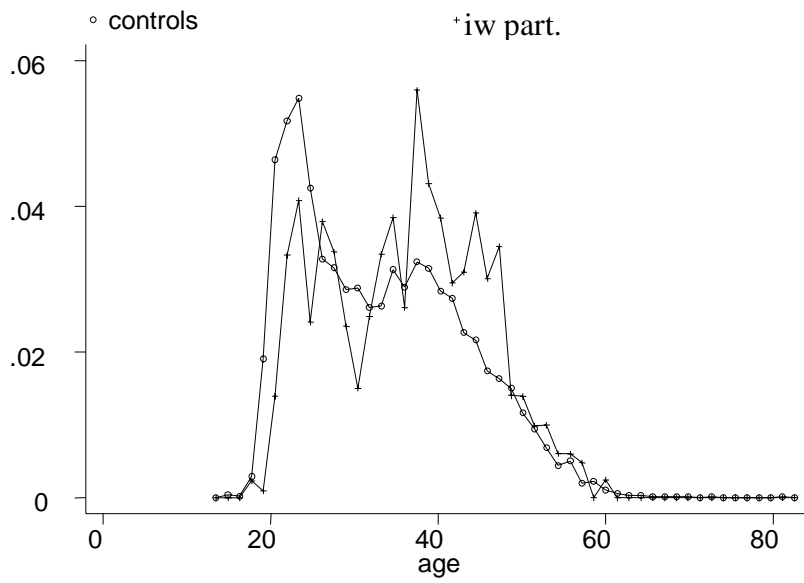


Figure 2. Distribution of age – Intervention Works

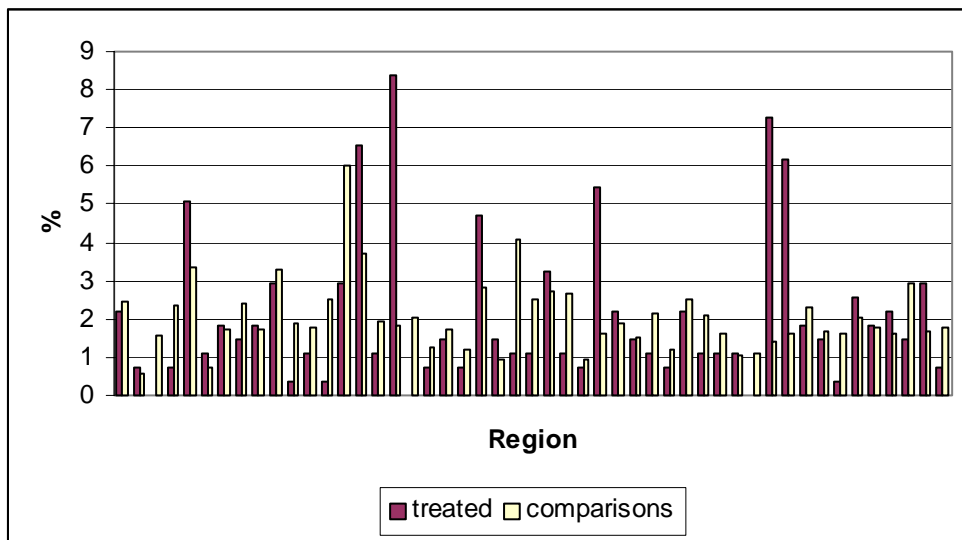
Sample A



Kernel density estimates of the relevant variable for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.5). Density estimates are not bound, their purpose is for illustration only.

Figure 3. Distribution of region – Intervention Works

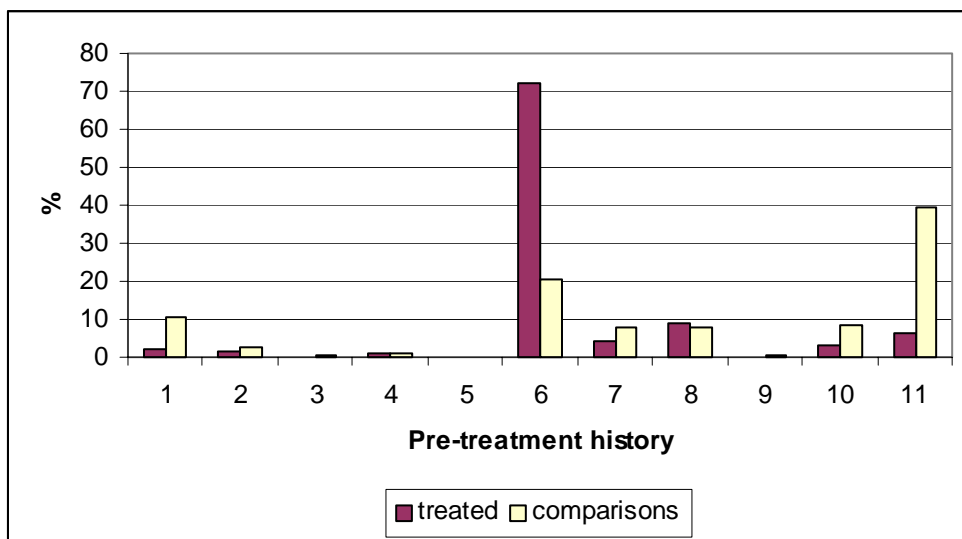
Sample A



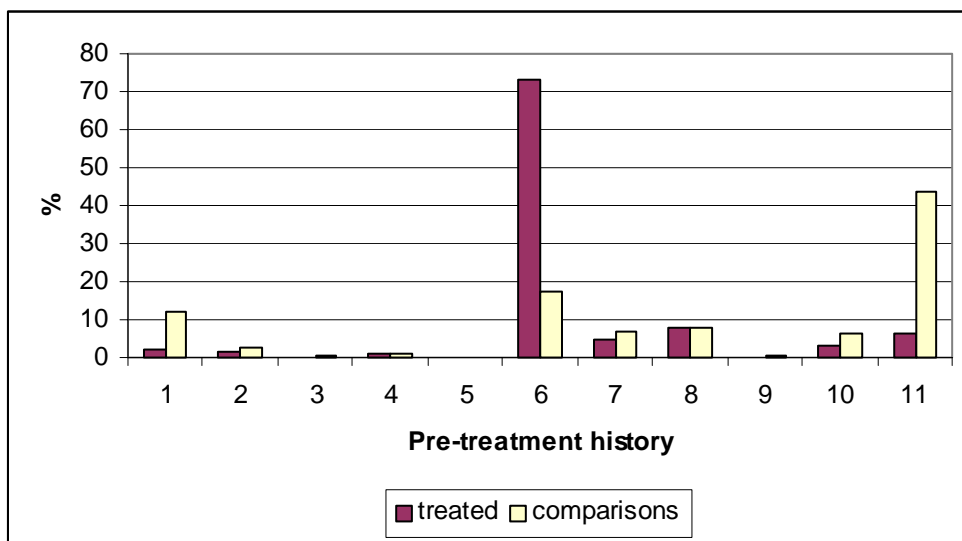
Region = 49 voivodships.

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

Sample A



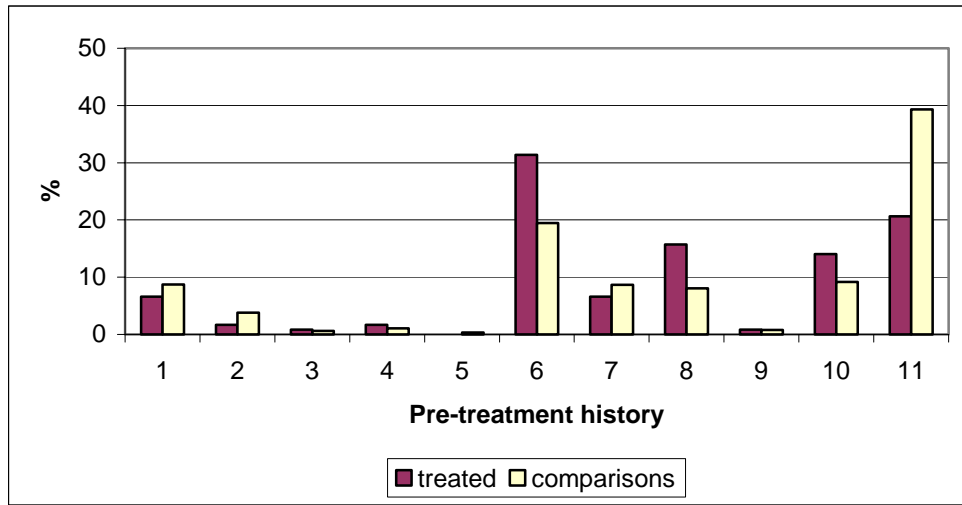
Sample B



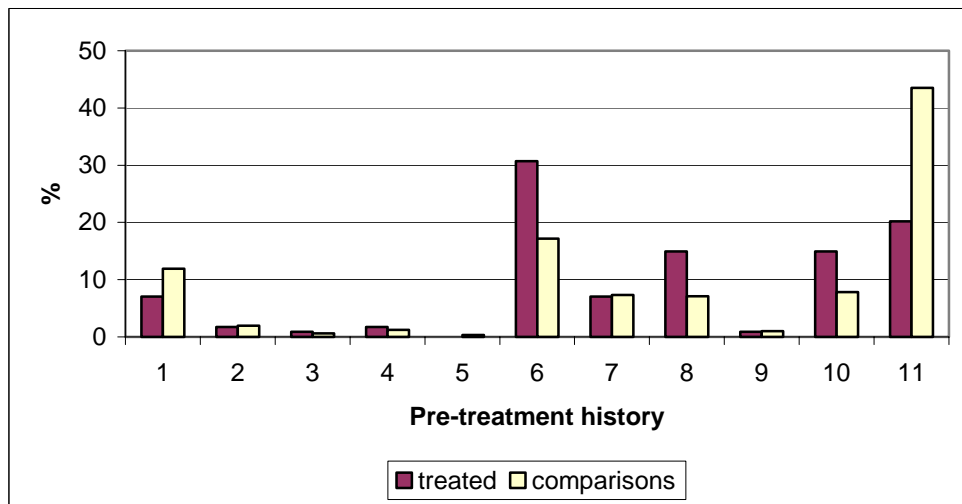
The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

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

Sample A



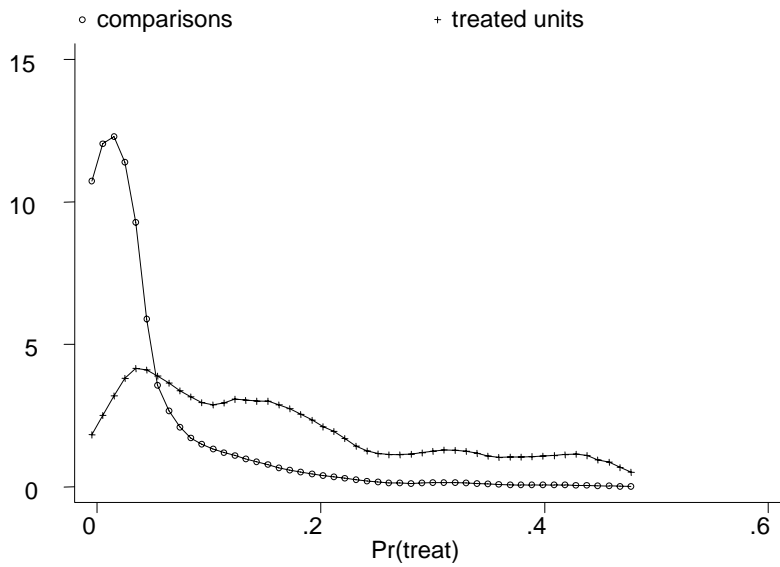
Sample B



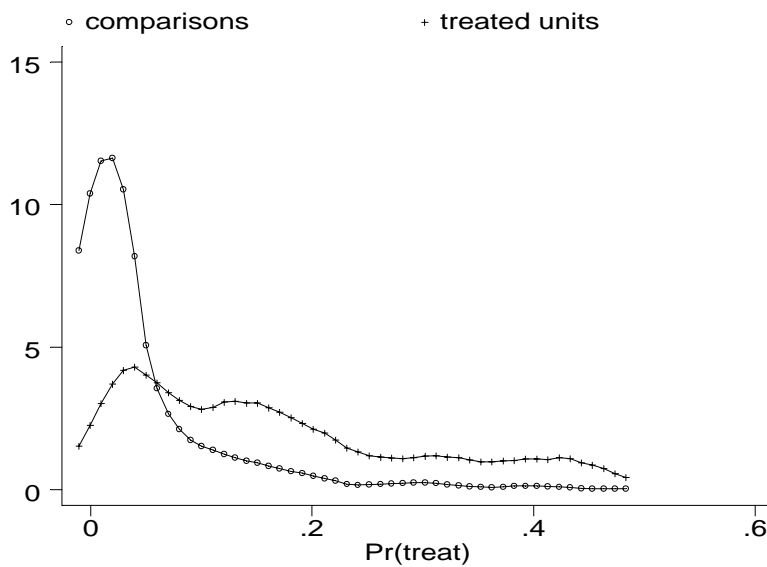
The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

Figure 6. Distribution of estimated propensity score by sample – Intervention Works

Sample A



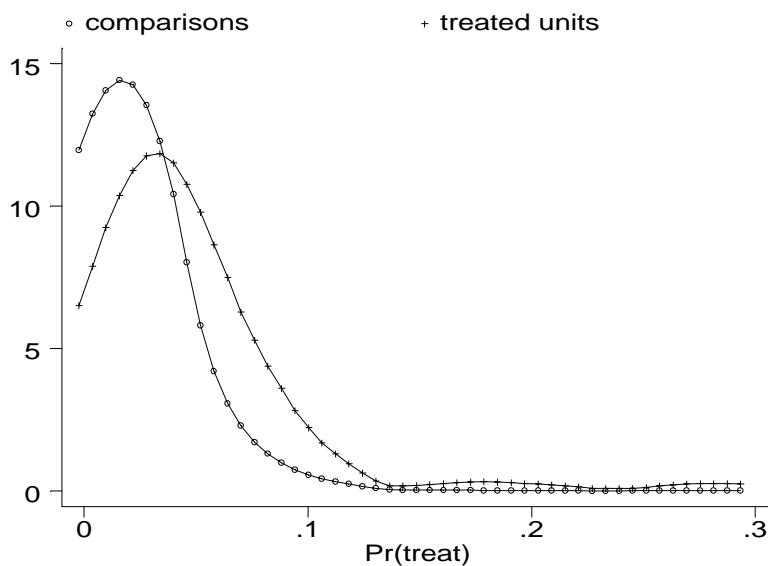
Sample B



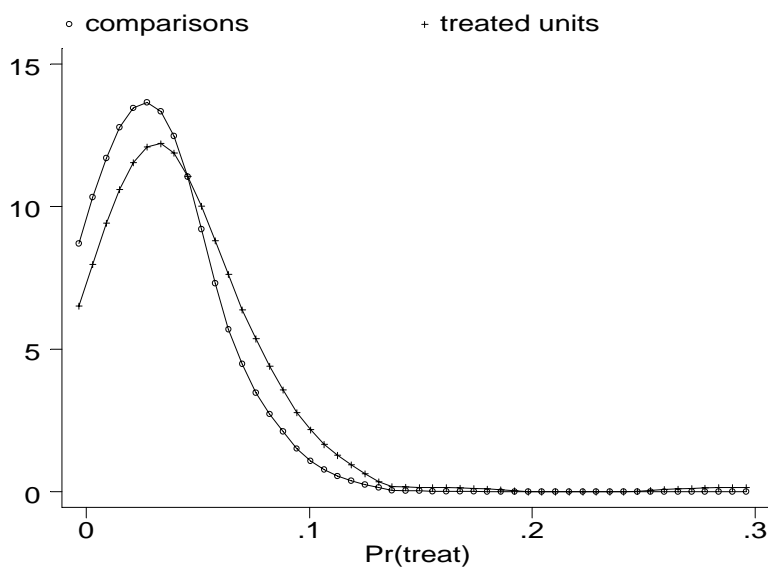
Kernel density estimates of the propensity score for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.02). Density estimates are not bound, their purpose is for illustration only. Y-axis denotes percentages.

Figure 7. Distribution of estimated propensity score by sample – Training

Sample A



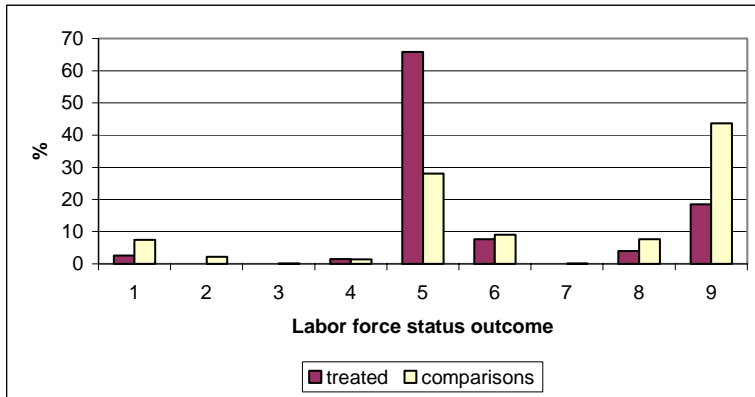
Sample B



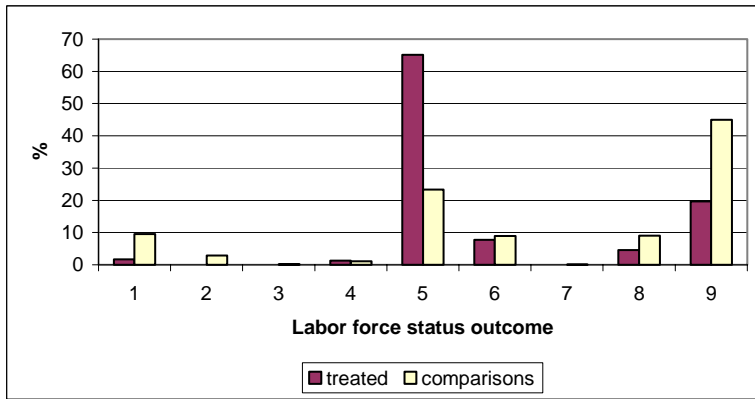
Kernel density estimates of the propensity score for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.02). Density estimates are not bound, their purpose is for illustration only. Y-axis denotes percentages.

Figure 8. Distribution of post-treatment labor market sequence by sample – Intervention Works

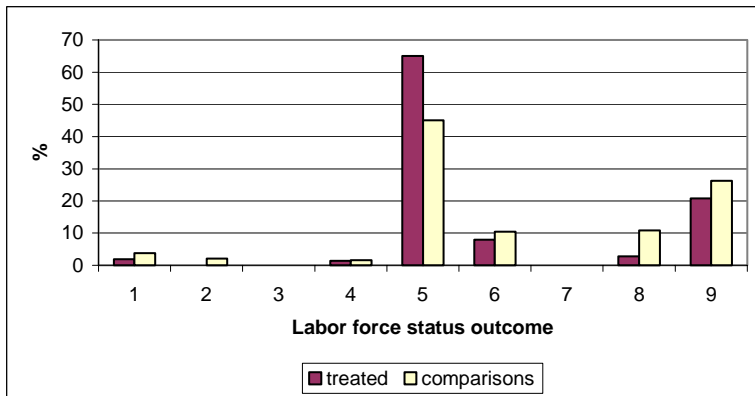
Sample A



Sample B



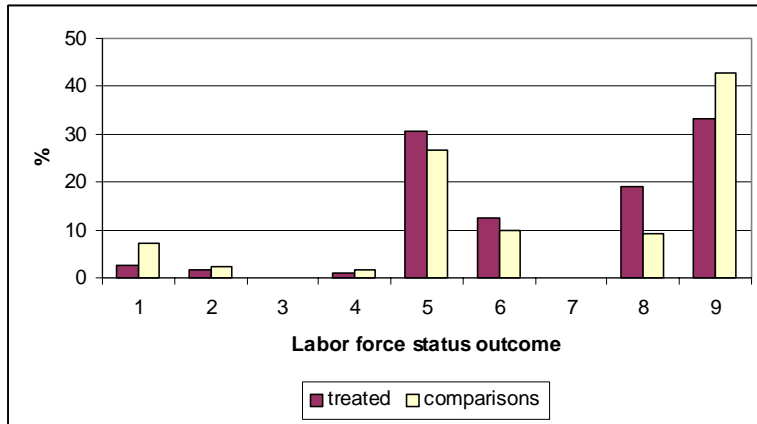
Sample C



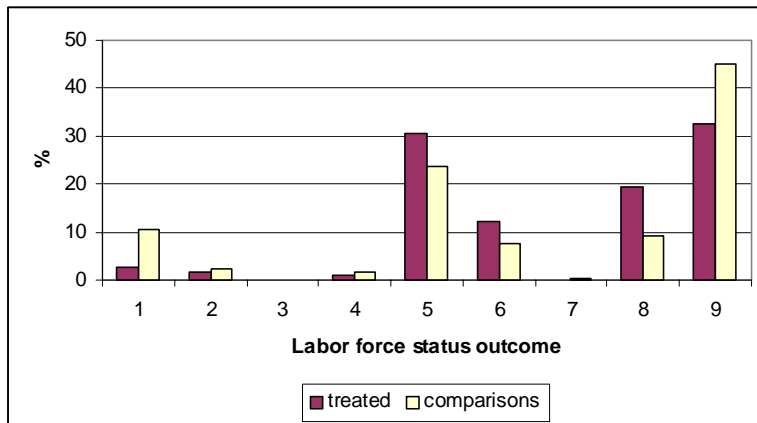
The 3³ possible labor force status sequences are classified into 9 categories (see text and Appendix A).

Figure 9. Distribution of post-treatment labor market sequence by sample – Training

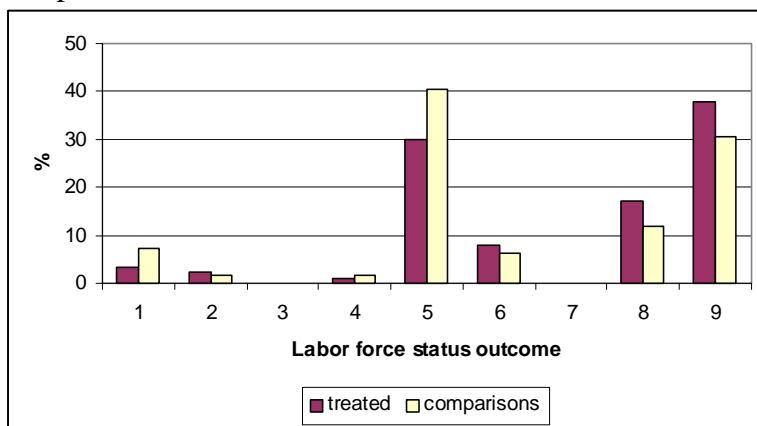
Sample A



Sample B



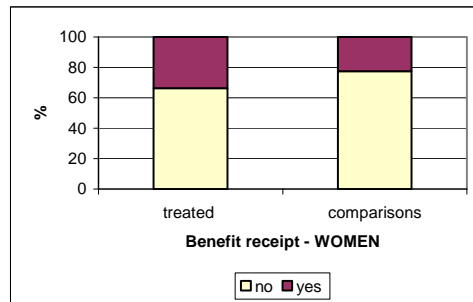
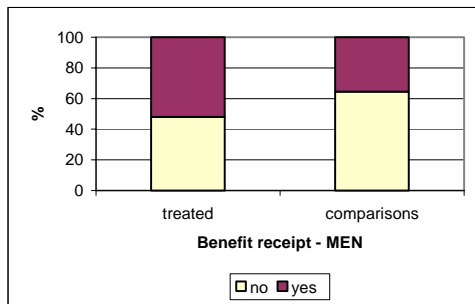
Sample C



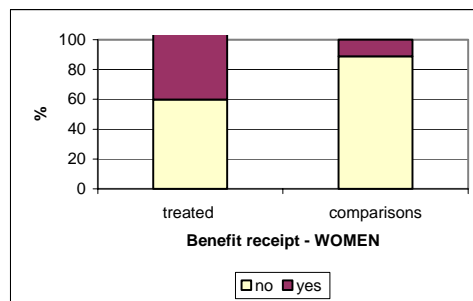
The 3³ possible labor force status sequences are classified into 9 categories (see text and Appendix A).

Figure 10. Distribution of benefit receipt by sex for sample C – Intervention Works

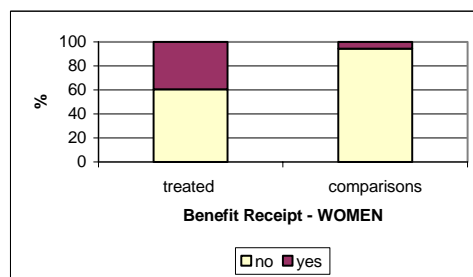
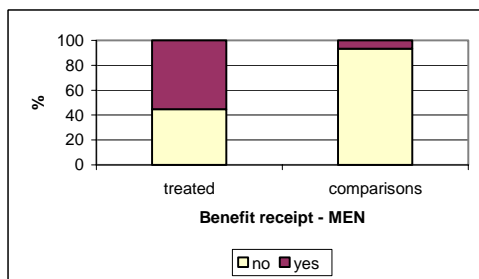
During 3 months BEFORE treatment:



During 3 months AFTER treatment:



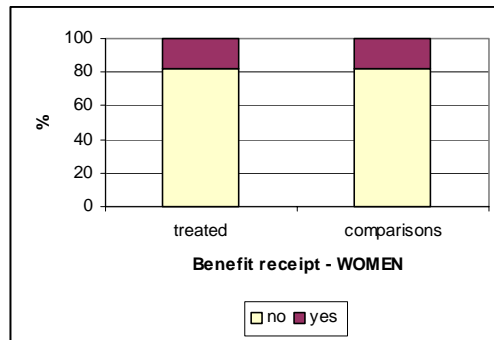
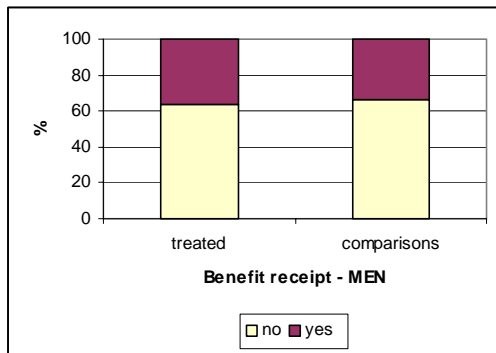
During 9 months AFTER treatment:



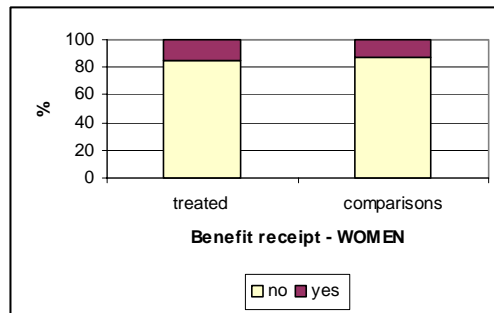
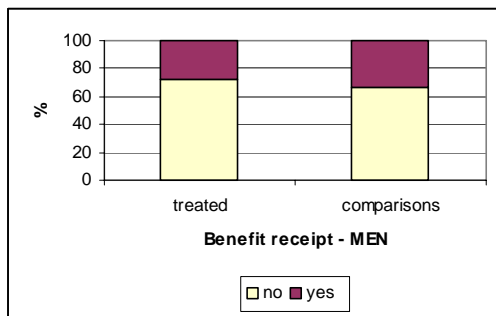
The upper panel indicates benefit receipt (= "yes") during at least two of the last three months preceding treatment. The middle panel indicates benefit receipt during at least two of the first three months succeeding treatment. The bottom panel indicates benefit receipt during at least two of the three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

Figure 11. Distribution of benefit receipt by sex for sample C – Training

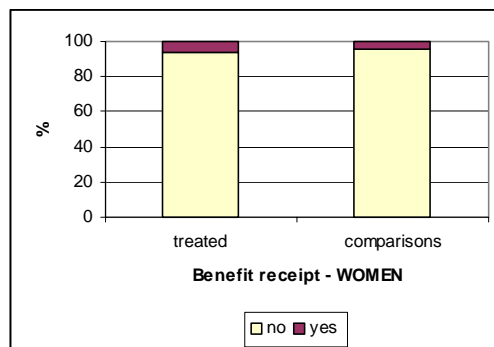
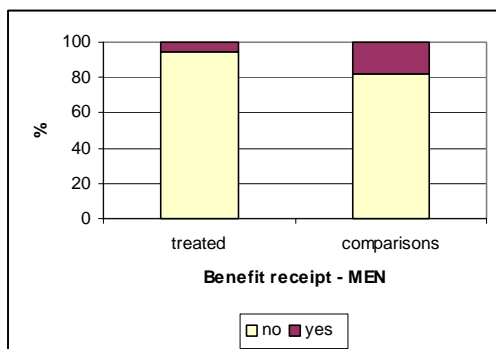
During 3 months BEFORE treatment:



During 3 months AFTER treatment:



During 9 months AFTER treatment:



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.