published in Annales d'Economie et de Statistique (Special Issue on Panel Data), 55-56, 97-128, 1999

The Effects of Enterprise-related Training in East Germany on Individual Employment and Earnings

Michael Lechner[*]

University of St. Gallen, Department of Economics

This version: March 1999

Abstract

The paper studies the returns from enterprise-related continuous vocational training on individual earnings, unemployment probabilities, and other labour market indicators in East Germany after unification. It attempts to solve the intrinsic identification problem of such evaluation problems nonparametrically by using restrictions ´produced´ by unification as well as by using very informative panel data (GSOEP, 1990-1994). The estimation is performed with nonparametric methods taking account of the panel structure. The results suggest that there are no effects with respect to employment and unemployment probabilities, but that there are large and positive earnings effects.

Keywords

Continuous vocational training, non-linear panel data, matching, East German labour markets.

JEL classification: C33, J24, J31, J60.

Michael Lechner
Swiss Institute for Applied Economic Reesearch (SIAW-HSG)
University of St. Gallen
Dufourstr. 48
CH-9000 St. Gallen, Switzerland
Email: Michael.Lechner@unisg.ch

[*] Financial support from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged. I thank the DIW for supplying the data of the German Socio Economic Panel. Furthermore, I thank Martin Eichler and Nadine Riede for carefully reading an early version of the manuscript and Klaus Kornmesser for competent help with the data. The paper has been presented at the 8th International Conference on Panel Data in Paris and at the Annual Meeting of the German Economic Association in Bern, 1997. I thank participants of both conferences and two anonymous referees of this journal for helpful comments and suggestions. All remaining errors are my own.

1 Introduction

This paper studies the returns from enterprise-related continuous vocational training (ERT) on individual earnings, unemployment probabilities, and other labour market indicators in East Germany after unification. Estimation of the effects of ERT mainly on earnings has received considerable attention in the literature in recent years. One reason is that this kind of training is considered important for continuously adapting the skills of the labour force to the requirements of technological change and hence for economic growth.

The case of East Germany is particularly interesting from an economic as well as an econometric point of view: The rapid transformation of the centrally planned economy of the former German Democratic Republic (GDR) to a West-German-style market economy requires a substantial adjustment of the skills of the labour force which has to be able to deal with new technologies and incentive systems. Despite the large flow of public money necessary to provide training for those of the labour force at risk of unemployment, for the majority of the labour force successful adjustment will crucially depend on the training efforts of the enterprises. ERT appears to be even more important, because there is empirical evidence that the publicly funded part and the off-the-job part of the early training efforts had no positive effects for the people participating in these programmes (see Lechner, 1996, 1999). Up to now, not much is known about the effectiveness of ERT in East Germany.

The methodological interest in this situation comes from the possibility to use the dramatic institutional changes due to unification with West Germany as well as a very informative panel data set available to identify the causal effects of training nonparametrically. Selectivity issues are an important issue for studies trying to answer causal questions about the returns of ERT. In typical microeconometric evaluations, outcomes measured for the sample undergoing the training are compared to outcome measures for a comparison group, that does not receive the training. In most social experiments such a group consists of individuals who apply for the programme, but are denied participation by randomization, for instance. However, such an experiment is usually not performed for ERT and was certainly not possible in the special circumstances of German unification. In a study not based on experimental data the researcher should find individuals who are identical to trainees regarding all *relevant* pre-training attributes except for not having obtained the training. Since typically such individuals cannot be easily identified, additional assumptions have to be invoked to adjust for their dissimilarity to avoid *sample selection biases*. Holland (1986) and Heckman and Hotz (1989) provide extensive and excellent discussions on these issues. Various econometric procedures are suggested in the literature to avoid such biases (e.g. Heckman and Robb, 1985).[1] However, for example LaLonde (1986) finds that the results are highly sensitive to different - equally plausible - stochastic assumptions made about the selection process. He concludes that social experiments are necessary to evaluate training programmes. Recently, Dehejia and Wahba (1995a, 1995b) re-evaluate the LaLonde (1986) data. By using nonparametric techniques that are similar to the ones used here, they come to more positive conclusions about the potential quality of inferences based on observational data. Here, one of the procedures suggested by Lechner (1999) that also takes account of the panel structure of the data is used to establish identification and to estimate with very few distributional or functional form assumptions. In that sense the results are expected to be robust.

In general there is a large number of evaluation results for ERT training programmes available, for example Groot, Hartog, and Oosterbeek (1994) and Lynch (1992, 1994).[2] Most of these papers are concerned with estimating the returns of ERT on wages. The corrections for selectivity are based on modelling the expectation of the outcome variable conditional on training participation and other factors. The exact type of model used depends, among other things, on whether panel data or only cross-sectional data is available. The results vary, but most studies find positive effects at least for subpopulations or specific types of training. So

[1] Chapter 1 in Bell, Orr, Blomquist, and Cain (1995) provides a more complete account of the development of the econometric evaluation literature.

[2] For further examples see the references in these papers.

far, there are only few econometric evaluations of ERT in East Germany. Fitzenberger and Prey (1996) evaluate the effects of ERT on employment as well as on earnings using different East German panel data. Their data is not as informative as the one used in this study, but has the advantage that it consists of a larger number of observations. Using standard type of assumptions for panel data random effects tobit and probit models, they specify the joint distribution of the outcome variables and the selection process to eliminate selection biases. They find positive effects of ERT on earnings, but not on employment probabilities.

All mentioned studies differ in many respects ranging from the database to the implementation of the evaluation, treatment of the selection problems, and the definition of the training itself. However, without using an explicit causality framework, they are all based on modelling the distributions of the outcome variables or error terms given certain observed or unobserved covariates.

The paper is structured as follows: The next section describes some features of the sample used for the empirical analysis and gives information on the type of training that is subject to the following evaluations. Additionally, so-called before-after comparisons are discussed. The causality concept used and its implementation for the current problem is discussed in Section 3. Section 4 contains the estimation results for the determinants of the probability of ERT participation as well as the evaluation results. Section 5 draws conclusions. Two appendices discuss further econometric aspects of the methods as well as results of specification tests.

2 The sample and some descriptive statistics

2.1 The GSOEP, sample selection and the definition of ERT

The sample used for the empirical analysis is drawn from the German Socio-Economic Panel (GSOEP), which is very similar to the US Panel Study of Income Dynamics (PSID). About 5000 households are interviewed each year beginning in 1984. A sample of just under 2000 East German households was added in 1990. The GSOEP is very rich in socio-demographic information, in particular concerning current and past employment status. For an English language description of the GSOEP see Wagner, Burkhauser, and Behringer (1993).

A very useful characteristic of this panel survey is the availability of monthly information between yearly interviews. This covers different employment and income states. The information is obtained by retrospective questions about what happened in a particular month of the previous year. Although this calendar does also contain information about vocational training, it is not possible to identify ERT. Therefore, the training information is taken from a special part concerned with continuous vocational training included in the 1993 survey.[3]

The definition of ERT used in all the empirical analysis is the following: The training takes place at least partly during regular working hours. Its aim is qualification other than retraining for a different occupation and familiarization with a new work place. Its duration is 16 hours or more in terms of full-time hours, or longer than one week in terms of overall duration. The purpose of this definition is to obtain a less heterogeneous group of trainees by excluding very short courses, off-the-job-training, retraining for a different occupation and familiarization with a new work place. The excluded types are different kinds of training with very heterogeneous objectives, probably with very different selection rules, and the possibility of receiving public funding. This definition does not exclude ERT-participants receiving some other kind of training before or after ERT-participation. It is worth emphasising that ERT is more formal than just learning-on-the-job. It consists of training courses that the employer allows or asks the trainee to attend during regular working hours. The implicit assumption is that in these cases the employer will at least indirectly bear part of the cost of ERT. It will become clear in the following that this indirect definition of ERT is partly motivated by the data available. Note that it is an important feature of this dataset that training begins at different points in time between mid 1990 and early 1993. The issue of

[3] For more information on relevant features of that survey see Lechner (1998).

dealing with different starting dates will be taken up in the discussion of a suitable estimator of the effects of ERT.

To be able to use the special survey as well as information concerning the employment status in the GDR, a sample of all individuals born between 1940 and 1970 who responded at least in the first four yearly interviews is selected. The upper age limit avoids the need of addressing early retirement issues. Since the population of interest is the labour force of the GDR, all selected individuals were working full-time just before unification. Additionally, individuals reporting severe medical conditions are not considered for obvious reasons. The entire labour market history before ERT (beginning in mid 1989) will be necessary to control for selection issues, hence it is required that all individuals answer the relevant questions of all four surveys during the first four years. Since the fifth survey (1994) is only used to measure post-ERT outcomes, such a requirement is not imposed for the final year (unbalanced panel).

2.2 Descriptive statistics

It has already be noted that the starting dates vary individually. They are almost evenly distributed over the years 1991 and 1992. In the second part of 1990 and in the first three months of 1993 only a few courses began. The former is probably due to the uncertainty following unification, and the latter is because the 1993 interviews (which contain the ERT information) started already in January. The ending dates are clustered more towards the second half of 1992.[4] The mean (standard deviation) of the ERT durations is about 1.8 (3.1) months, and its median is about 0.6 months of full time training. 21% of the courses have a duration of one week or less of full time training, 47% between one week and one month,

[4] Combining the information on raw durations and intensity (hours per week) of ERT, the ´durations´ are expressed in terms of weeks or months of full time training (38 h per week, 4.3 weeks per month).

16% between one month and three months, and the remaining 16% have a duration of more than three months.[5]

Let us consider the employment status of ERT participants before and after ERT. Figure 1 shows the unemployment rate. ERT participants experience very little unemployment before ERT participation. After ERT their unemployment rate increases from 3% to more than 9%. A similar picture appears when using the share of full time employment instead. It would be unjustified to conclude from these figures that ERT increases unemployment and reduces full time employment. Quite contrary, these figures clearly show the limitations of using so-called before - after comparisons for evaluation, particularly in non-stationary economies. However, the effects of the contracting economy (that is the effect of calendar time) can be eliminated by choosing a comparison group in a random fashion. When doing so, the above conclusions are reversed and it appears that at least in the short run ERT has positive employment effects.[6] However, as will be shown in section 3 these conclusions are flawed as well, because the comparison group needs to be chosen in a more sophisticated way.

< ------------------------------------ *Figures 1 and 2 about here* ------------------------------ >

Figure 2 shows a similar plot for the real gross earnings' variable. Earnings are measured only for the month preceding the yearly interview. Furthermore, it does not capture bonuses, etc., paid only at the end of each year. The deflator used is the cost-of-living price index. Therefore, the sharp increase of average earnings may merely reflect the divergence of wage growth and cost of living. In this figure earnings for non-workers are coded as unemployment benefits, but the same shape of the curve emerges when earnings for non-workers are coded as zeros. Comparing Figure 2 with the earnings' difference of ERT participants and randomly chosen non-participants shows that the latter have lower mean earnings before as well as after

[5] 8% of the observations are possibly right censored. However, the measurement of the ending dates is not exact. To avoid classifying labour market outcomes during training as post-training outcomes, the maximum

ERT.[7] This does not imply a positive effect of ERT, but merely is another indication that ERT participants are a very selective group with lower unemployment probabilities and higher earnings' capacity. Correcting for these selection effects will be important to obtain reliable estimates of the effects of ERT only.

< ------------------------------------ Table 1 about here ------------------------------------ >

Considering the marginal distributions of other selected socio-economic variables given in Table 1 again shows that ERT participants are a very selective group. Women are underrepresented, and there is a correlation between ERT participation and level of education as well as job position. Those having obtained a higher degree of education and / or a higher job position are far more likely to be observed in ERT. Hence, they are also better paid. Furthermore, those working in managerial, scientific, technical and medical occupations are more likely to be observed in ERT than individuals working on the production floor. Considering the industrial sectors, agriculture, the light industry, and trade are on the negative side with respect to ERT participation, whereas energy, water, and the aggregated group of other services (non-profit, banks, insurance, government, legal, personal services, cleaning, waste disposal, hotels, restaurants) are on the positive side. The worries about job security and expectations about redundancies in the firms employing the individuals show that ERT participants are more optimistic about their firm as well as about the security of their job.[8]

duration is assumed for the computation of the ending dates. Therefore, a substantial, but unknown part of these 8% is not censored at all.

[6] See Lechner (1998).

[7] Note that Ashenfelter´s (1978) dip in earnings is absent or even reversed. This is not surprising because *his* famous dip resulted by selecting unemployed individuals into training (see also Heckman and Smith, 1995). The previous figure shows that in this study the selection process works just in the opposite direction.

[8] Lechner (1998) presents additional descriptive information on ERT based on GSOEP data and other studies.

3 Causality, identification, and empirical implementation

3.1 Causality and identification

The empirical analysis attempts to answer questions like "What is the average gain for ERT participants compared to the hypothetical state of nonparticipation?" (the average treatment effect on the treated). The question refers to potential outcomes. The underlying notion of causality requires the researcher to determine whether participation or nonparticipation in ERT affects the respective outcomes, such as earnings or employment status. This is different from asking whether there is an empirical association, typically related to some kind of correlation, between ERT and the outcome.[9]

The framework serving as guideline for the empirical analysis is the potential-outcome approach to causality suggested by Rubin (1974). This idea of causality is inspired by the set-up of experiments in science. Main building blocks for the notation are *units* (here: individuals), *treatment* (participating in ERT or not) and potential *outcomes*, that are also called *responses* (earnings, labour market states, etc.). Y^t and Y^n denote the outcomes (*t* denotes treatment, *n* no treatment).[10] Additionally, denote variables unaffected by treatments - called *attributes* by Holland (1986) - by *X*. Attributes are exogenous in the sense that their potential values for the different treatment states coincide ($X=X^t=X^n$). Also, a binary *assignment* indicator *S* is defined determining whether unit *n* gets the treatment ($S = 1$) or not ($S = 0$). When participating in ERT the observable outcome variable (Y) is Y^t, and Y^n, otherwise.

The average causal effect of ERT (θ^0) for participants is defined in equation (1):

$$\theta^0 := E(Y^t - Y^n \mid S=1) = E(Y^t \mid S=1) - E(Y^n \mid S=1)\,. \qquad (1)$$

[9] See Holland (1986) and Sobel (1994) for an extensive discussion of concepts of causality in statistics, econometrics, and other fields.

[10] As a notational convention big letters indicate quantities of the population or of members of the population and small letters denote the respective quantities in the sample. The units of the sample ($i=1,...,N$) are supposed to come from *N* independent draws in this population.

E(·|S=1) denotes the mean in the population of all units who participate in training. To draw inferences in subpopulations of *S=1*, defined by attributes in *X*, the respective expressions are changed in an obvious way. Note that inference is about the average effect. An assumption about the kind of variation of the treatment effect in the population (popular example: same effect for all participants) is not necessary.

θ^0 cannot be identified without further assumptions, because the sample analogue of $E(Y^n \mid S=1)$ - the mean of y_i^n for participants $(s_i = 1)$ - is unobservable. Much of the literature on causal models in statistics and selectivity models in econometrics is devoted to find reasonable identifying assumptions to predict the unobserved expected nontreatment outcomes of the treated population by using the observable nontreatment outcomes of the untreated $(y_i^n, s_i = 0)$ in different ways.

If there is random assignment as in a suitably designed experiment, then the potential outcomes are independent from the assignment mechanism and $E(Y^n \mid S=1) = E(Y^n \mid S=0)$. Thus the untreated could be used as the control group, because the expectation of their observable outcome equals $E(Y^n \mid S=1)$. Given a large enough sample, the corresponding sample moments converge towards these population moments under standard regularity conditions. However, the assumption of random assignment is not satisfied in this study, because there are several variables influencing assignment as well as outcomes (see above).

Using the law of iterated expectations to rewrite the crucial part of equation (1) as:

$$E(Y^n \mid S=1) = E[E(Y^n \mid S=1, X=x) \mid S=1], \tag{2}$$

it becomes clear that an assumption leading to $E(Y^n \mid S=1, X=x) = E(Y^n \mid S=0, X=x)$ is sufficient to identify $E(Y^n \mid S=1)$, since $E[E(Y^n \mid S=0, X=x) \mid S=1]$ could then be estimated by standard methods (note however that the outer expectation operator is with respect to the distribution of *X* in the population of participants). Rubin (1977) proposed such an assumption, called random assignment conditional on a covariate. The assumption is that the assignment is independent of the potential non-treatment outcome conditional on the value of a covariate or attribute (CIA). The following sections will show that this restriction is reasonable in the context under investigation. The task will be to identify and observe all variables that could be correlated with assignment and potential nontreatment outcomes. This implies that there is no variable left out that influences nontreatment outcomes as well as assignment given a fixed value of the relevant attributes.[11] Since in our case, such attributes will necessarily measure also the employment history prior to ERT, panel data become imperative to identify the effects of ERT in our context.

Rosenbaum and Rubin (1983, RR) show that if CIA is valid the estimation problem simplifies. Let *P(x) = P(S=1|X=x)* denote the nontrivial participation probability *(0 < P(x) < 1)* conditional on a vector of characteristics *x*. *P(x)* is called the propensity score. Furthermore, let *b(x)* be a function of attributes such that *P[S=1|b(x)] = P(x)*, or in the words of RR, the balancing score *b(x)* is at least as 'fine' as the propensity score. RR show that if the potential outcomes are independent of the assignment conditional on *X*, they are also independent of the assignment conditional on *b(X)*, hence:

$$E[Y^n \mid S = 1, b(X) = b(x)] = E[Y^n \mid S = 0, b(X) = b(x)]\,, \tag{3}$$

and $E(Y^n \mid S = 1) = E\{E[Y^n \mid S = 0, b(X) = b(x)] \mid S = 1\}$ can be used for estimation. The advantage of this property is the reduction of dimension of the (nonparametric) estimation problem. However, the probability of assignment - and consequently any dimension reducing balancing score - is unknown and has to be estimated. This estimation may also lead to a better understanding of the assignment process itself.

[11] In the language of regression-type approaches such a variable leads to simultaneity bias.

3.2 Identification, economic theory, and information in the sample

A detailed economic analysis of the participation process into ERT will identify age, expected labour market prospects, actual employment status, and other socio-economic and firm characteristics as major factors that could potentially influence the ERT participation (see Lechner, 1998). Before going into more details about the groups of variables used in the empirical analysis, I will discuss more fundamental issues concerning the admissibility of variables in the conditioning set. Additionally, I will state two assumptions that are very important in that respect for the particular situation in East Germany after unification, because they make CIA a powerful and justifiable assumption in this specific context.

To use the language from the previous section, a variable *W* is allowable for the set of attributes *X* if their potential values do not depend on the treatment status. Obvious candidates for *X* are time constant variables or variables dated prior to ERT. However, some of the latter variables may be problematic. For example, consider the case when an employer and an employee explicitly or implicitly agree on a cost sharing scheme for ERT that reduces the earnings of the employee by a given amount in the year before ERT starts. Clearly, pre-ERT earnings can no longer be an attribute. Instead it is an outcome, because its decline before ERT is caused by ERT itself. The same is true for other employment- related variables or expectations about the career. This situation is very unsatisfying, because on the one hand the closer the information is to the start of ERT the more informative it should be as an attribute, but on the other hand the more likely it is to make CIA inplausible (since W is *endogenous* in that sense). Since there is no information on the actual date of the *decision* to participate in ERT, and since the arrangement will certainly vary from firm to firm, there is no easy way out. However, the following assumptions will probably reduce the problem by a substantial amount.

The first assumption is that the complete switch from a centrally planned economy to a market economy in mid 1990, accompanied by a completely new incentive system, invalidates any long term plans that connect past employment behaviour to ERT participation. It was generally impossible to predict the impact and timing the change of the system would have. Even when it was partly correctly foreseen, it was generally impossible to adjust behaviour adequately in the old system. This is true for workers as well as for firms. This assumption is further supported by the fact that almost all firms changed ownership at some point in time after unification. This assumption allows the use of all pre-unification variables as components of *X*.

An additional assumption invoked is related to the condition of the labour market in the rapidly contracting East German post-unification economy. The labour market is characterized by rapidly and continuously rising unemployment. Furthermore, only about 10% of those working full-time in mid 1990 were sure that they would not lose their job within the next two years. It is assumed that no individual - having only slim chances of getting rehired once being unemployed - will voluntarily give up employment (or become self-employed) to get easier or cheaper access to ERT later. Note that this does not preclude a change of employer for that reason as long as job search does not result in a spell of nonemployment between the two jobs. This assumption allows the use of monthly pre-training information on full-time employment, involuntary short-time work, and unemployment as components of *X*.

The groups of variables that are used in the empirical analysis to approximate and describe the above-mentioned four broad categories of factors are age, sex, marital status, educational degrees as well as regional indicators. Features of the pre-unification position in the labour market are captured by many indicators including wages, profession, job position, employer characteristics such as firm size or industrial sector. Individual future expectations are described by individual pre-unification predictions about what might happen in the next two years regarding job security, a change in the job position or profession, and a subjective conjecture whether it would be easy to find a new job or not.[12] Furthermore, monthly employment status information, as mentioned before, is available from July 1989 to December 1993.

Having discussed potentially important factors and variables available for the empirical analysis, the question is whether any important group of variables is missing. One such group can be described as motivation, ability, and social contacts. I approximate such attributes by the subjective desirability of selected attitudes in society in 1990, like 'performing own duties', 'achievements at work', and 'increasing own wealth', together with the accomplishment of voluntary services in social organizations and memberships in unions and professional associations before unification, as well as schooling degrees and professional achievements. Additional variables indicate that the individual is not enjoying the job, that income is very important for the subjective well-being, that the individual is very confused by the new circumstances, and optimistic and pessimistic views of general future developments. Another issue is the discount rate implicitly used to calculate present values of future income streams. I assume that controlling for factors that have already been decided by using the individual discount rate, such as schooling and professional education, will be sufficient.

In conclusion, it seems safe to assume that these missing factors (conditional on all the other observable variables) play only a minor role. However, the endogeneity problem of some pre-ERT employment variables remains. Although endogeneity of all pre-ERT employment variables does not seem to be very likely, the empirical analysis will check the sensitivity of the results in that respect.

[12] Details of the particular variables - mostly indicators - as well as their means and standard errors in the training and comparison groups are available on request (see footnote 5).

3.3 Empirical implementation

3.3.1 Estimators

This section summarizes the estimation methods as suggested by Lechner (1999) and applied in Lechner (1996, 1999).[13] The considerations in the previous sections suggest to estimate the causal effects with nonparametric methods in order to avoid the consequences of potential misspecifications. For notational convenience assume that observations in the sample are ordered such that the first N^t observations receive ERT, and the remaining $(N\text{-}N^t)$ observations do not. The following nonparametric regression estimator is an obvious choice:

$$\hat{\theta}_N = \hat{E}(Y^t - Y^n \mid S = 1) = \frac{1}{N^t}\sum_{i=1}^{N^t} y_i - \frac{1}{N^t}\sum_{i=1}^{N^t} \hat{g}^n[b(x_i)]. \tag{4}$$

$\hat{\theta}_N$ denotes the consistent estimate of the causal effects and $\frac{1}{N^t}\sum_{n=1}^{N^t}\hat{g}^n[b(x_n)]$ denotes a consistent estimate of $E(Y^n \mid S = 1)$. Consistency is satisfied under standard conditions, if $\hat{g}^n[b(x_n)]$ is asymptotically unbiased for $E[Y^n \mid S = 0, b(X) = b(x_n)]$. Nonparametric regression could be used to provide such an estimate. However, the balancing score most useful in this particular evaluation study necessarily has a high dimension (see below). Therefore, and given the size of the available sample, nonparametric regressions are subject to the typical *curse* of dimensionality.

For these reasons a matching approach (e.g. Rosenbaum and Rubin, 1983, 1985) is used (see Appendix A.1). The idea is to find for every treated observation a single comparison observation that is as close as possible to it in terms of a balancing score. When an identical comparison observation is found, the estimation of the causal effect is unbiased.[14] In cases of

[13] The interested reader is referred to Lechner (1995) for more details on the estimation methods.

[14] Compared to the nonparametric regression described above, there is an asymptotic efficiency loss because observation i ($i \leq N^t$) and its closest neighbour in the *comparison* population - instead of the many possible close neighbours - are used to compute $\hat{g}^n[b(x_i)]$.

'mismatches', it is often plausible to assume that local regressions on these differences remove the bias (see Appendix A.2 for details).

Define the differences in matched pairs in the sample as $\Delta y_i = y_i^t - y_j^n$, $\Delta b(x_i) = b(x_i^t) - b(x_j^n)$, $i = 1,...,N^t$, where y_j^n and x_j^n denote values of an observation from the pool of individuals not participating in ERT (comparisons) that is matched to the ERT observation *i*. The estimate of the average causal effect and the respective standard error are computed as:

$$\hat{\theta}_{N^t} = \frac{1}{N^t}\sum_{i=1}^{N^t}\Delta y_i, \qquad Var(\hat{\theta}_{N^t}) = \frac{1}{N^t}(S_{y^t}^2 + S_{y^n}^2). \tag{5}$$

$S_{y^t}^2$ and $S_{y^n}^2$ denote the square of the empirical deviation of *Y* in the ERT sample and in the sample matched to the ERT-sample, respectively.[15] As mentioned in the previous section, when a perfect match is achieved, i.e. $\Delta b(x_i) = 0$, $i = 1,...,N^t$, these estimates are unbiased. In a sufficiently large sample, the normal distribution can be used to perform tests and to compute confidence intervals.

Equation (5) gives the principal nonparametric estimate of the causal effect to be refined in the following to take account of time before and after ERT. Denote by N_τ^t, $\tau \in \{...,-3,-2,-1,1,2,3,...\}$ the number of pairs observed at any distance to ERT.[16] Let $\iota_\tau(i)$ = 1 if observation *i* is observed at distance τ. The observability of an observation in a particular post-ERT distance on the redefined time scale depends only on the ending date of ERT (unbalanced panel). I assume that they are independent random variables.[17] The refined estimators are defined as:

[15] The variance estimate exploits the fact that the matching algorithm never chooses an observation twice.

[16] Note that ERT starting dates vary individually. Now we switch from calendar time to time relative to the beginning and the end of ERT. The whole ERT period that varies also individually is denoted as period 0.

[17] Two checks are performed with respect to this assumption. First of all, the ending dates (months) are regressed on $(1,p(v),p(v)^2,p(v)^3,m)$ (*v,m* is explained in section 3.3.2). None of the variables, except the constant, is significant at the 5% level. The adjusted R^2 is 0.07 (*N*=185). Secondly, the sample is split according to different ending dates, but the qualitative results do not change. Therefore, there is no evidence from the data that the independence assumption (typically used in unbalanced panels) is suspect.

$$\hat{\theta}_{N_\tau^t} = \frac{1}{N_\tau^t}\sum_{i_\tau=1}^{N^t}\iota_\tau(i)\Delta y_{i,\tau}\ , \qquad \tau \in \{\ldots,-3,-2,-1,1,2,3,\ldots\}\ ; \tag{6}$$

$$\hat{\theta}_{N_\tau^t}^{T} = \frac{1}{N_\tau^t}\sum_{i_\tau=1}^{N^t}\sum_{T=1}^{\tau}\iota_\tau(i)\Delta y_{i,\tau}\ , \qquad \tau \in \{1,2,3,\ldots\}\ . \tag{7}$$

The variances are computed appropriately. When τ is negative, then $\hat{\theta}_{N_\tau^t}$ denotes the mismatch in period τ before ERT, otherwise it denotes the effect of training in period τ after ERT. $\hat{\theta}_{N_\tau^t}^{T}$ indicates the accumulated effect τ periods after ERT. Note again that the individual training effects are allowed to vary freely in the population.

3.3.2 Balancing score and partial propensity score

The estimation of the propensity score is not straightforward, because there are potentially important variables - monthly pre-training employment status and yearly pre-training self-employment for example - that are related to the distance (months or years) of the beginning of ERT.[18] Since these dates differ across ERT participants, they are not clearly defined for the comparison group. Consequently, a particular form of balancing score that is different from the prototypical propensity score has to be used.

Partition the vector of observed attributes in two groups such that $X = (V, M)$, and suppose that $P(S=1|X=x) =\ P(x) =\ P[V\beta^0 + f(M,U) > 0|V=v, M=m]$. U denotes attributes - not included in X - that are independent of the potential outcomes, but influence ERT participation. V contains time invariant attributes. β^0 is a fixed parameter vector. M denotes time variant pre-training variables. If the potential outcomes are independent of S conditional on $P(X)$ = $P(x)$, then they are also independent of S conditional on ($V\beta^0 = v\beta^0,\ \ M = m$), because $(v\beta^0, m)$ is a balancing score. The use of $v\beta^0$ instead of v can still lead to a large reduction of the dimension of the conditioning set.

$v_i\beta^0$, $i=1,...,N$, is estimated by maximum likelihood using a a binary probit model. The basic condition for the consistent estimation of $v_i\beta^0$ up to scale is that the conditional expectation of the dependent variable is correctly specified:

$$P(S=1 \mid V\beta^0 = v_i\beta^0) = \Phi(v_i\beta^0), \qquad i=1,...,N. \tag{8}$$

$\Phi(v_i\beta^0)$ denotes the cumulative distribution function of the standard normal distribution evaluated at $v_i\beta^0$. The first of two sufficient conditions for equation (8) to hold is that the propensity score has the additive form $P(x) = \ P[V\beta^0 + f(M,U) > 0 | V = v, M = m]$. This assumption is not so restrictive, because *V* may contain flexible functional forms for the attributes, such as polynomials or interaction terms. The crucial assumption is that:

$$[f(M,U)] [V\beta^0 = v\beta^0] \sim N(0,1). \tag{9}$$

$N(0,1)$ denotes the standard normal distribution. The crucial assumptions are normality and mutual independence of *f(M,U)* and $V\beta^0$. They are tested with several specification tests.

4 Results

4.1 Participation in ERT: partial propensity score

The main purpose of the estimation of $v_i\beta^0$ is to obtain *good* predictions for the partial propensity score to find comparison observations that are similar to ERT observations. Therefore, in several cases without or almost without any treated observations in certain cells of *v*, observations in these cells are deleted from the sample. This leads to a loss of 11 ERT and 259 comparison observations. Hence one should note that most of these variables would have appeared with a negative sign in the estimations. The second remark concerns the splitting of the sample in three separate parts according to gender and according to job

[18] This section summerizes one of the proposals of Lechner (1999).

position (men only): The heteroscedasticity as well as the information matrix tests reject very strongly the hypothesis that the conditional model is the same for these three groups. This remains true even after including several interaction terms of gender and job positions. Note that the total sample sizes (men with highest job position: 150, other men: 524, women: 503) as well as the share of treated observations (43%, 18%, 15%) differ substantially in the subsamples. The conditional effects of the variables used for splitting the sample can therefore not be estimated.[19] The results of the estimation are given in Table 2.

< ------------------------------------ *Table 2 about here* ---

>

The descriptive statistics and the estimation results given in Table 2 indicate that individuals with higher *job positions* and higher *degree*s are more likely to participate in ERT. Furthermore, working on the *production* floor is related to lower ERT participation. There are no significant differences with respect to industrial sectors for men with the highest job positions, but for other men working in the sector *energy and water* is significantly positively correlated with ERT and working in *trade* is significantly negatively related to ERT. For women negative relations with *agriculture* and positive relations with the sector *other services* (non-profit organizations, banks, insurance, government, legal, personal services, cleaning, waste disposal, hotels, restaurants) appear. Additionally, the expected impact of *firm size* appears also significantly: small firms have ceteris paribus less ERT participation. The threat of unemployment that is approximated by several variables has an ambiguous effect in all subsamples. Finally, there are some regional effects as well as a negative age effect for men in lower job positions.

[19] A comparison of the constant terms does not give such an estimate, because identification is only up to scale. The scales may differ across subsamples. Splitting subsamples is consistent with the following generalization of the assumptions mentioned in the previous section. Denote the conditional error variance by $\sigma_f^2(v) = Var[f(M,U)|V = v]$. Then the participation probability (propensity score) can be rewritten as $P[V\beta^0 / \sigma_f(V) + \; f(M,U) / \sigma_f(V) > 0|V = v, M = m]$. The respective balancing score is $[v\beta^0 / \sigma_f(v), m]$. Estimation in the splitted samples estimates $[\beta^0 / \sigma_f(v)]$ consistently if $[f(M,U) / \sigma_f(V)|V = v] \sim \; N(0,1)$.

The other variables mentioned before are considered as well by means of a score test against omitted variables, but none of them appeared to be missing in the partial propensity score. For women, the specification tests do not provide any evidence against the chosen specification. For men some rejections occur for the heteroscedasticity tests and the information matrix tests. However, for neither of the samples do the normality tests reject. The same is true for the information matrix test based on all possible indicators. This test is known to be a powerful omnibus tests. Its non-rejection gives some confidence in the overall fit of the model. The complete results are contained in Appendix B. Nevertheless, some components of *v* are included in the balancing score in addition to the estimated partial propensity score to reduce the impact of any misspecification resulting from a possibly inconsistent estimation of β^0.

4.2 Participation in ERT: similarity of ERT participants and matched comparison groups

A requirement for a bias removing matching is an overlap of the distributions of the conditioning variables in both subsamples. For a very important conditioning variable, $v\hat{\beta}$, the mass of the distribution of the comparisons is to the left of the treated, but there is overlap for a large part of the distribution of ERT participants. However, there is a lack of overlap in the right tail of the distributions. Hence, it is unlikely that matching alone is successful in removing all bias, so that the local econometric adjustment procedures proposed in Lechner (1999) are also used in the empirical analysis (see Appendix A.2).

< ------------------------------------- *Table 3 about here* --- >

Table 3 confirms these considerations. The table gives descriptive statistics of selected variables for ERT participants and two matched comparison groups. The first comparison group (*comparison 1*) does not include the latest post-unification-pre-training information about job and employer characteristics as matching variables (*m*), whereas the second *comparison* group (*comparison 2*) does include that information. A comparison of the mean of the partial propensity score shows that the match is indeed not perfect. The table reveals that the largest problem is a lack of sufficiently well educated individuals who are also similar to the ERT participants with respect to other relevant characteristics. Comparing the two different *comparison* groups no substantial differences appear, although for most variables *comparison* group 1 appears to be closer to ERT than *comparison* group 2. This is expected, since group 2 is constructed by taking more matching variables into account than for group 1. Additional information on the match-quality with respect to time varying pre-ERT variables is given below.[20]

4.3 Evaluation results

4.3.1 Introductory remarks

This paper is particularly interested in the effects of ERT on post-training changes in the actual labour market status. It is due to the nature of the data and the circumstances (German unification in **1990**) that at the time this paper was written no long run effects of ERT could possibly be discovered.

The following outcomes are measured on a monthly basis by way of the retrospective employment calendar: involuntary short-time work, registered as being unemployed, and full-time employment. Another variable capturing characteristics of the actual labour market status - measured once a year - is gross monthly earnings. For those being employed, it is defined as the gross monthly earnings in the month before the interview. For those not being employed, either imputed unemployment benefits or social assistance - whichever is higher - or zeros are used instead. Although the focus is on this group of variables, there is also information available about the individual labour market prospects. The respective variables

[20] When comparing different matching algorithms, that differ by the choice of *m*-variables, it is found that it is

are measured once a year as individual expectations or worries. They include expectations whether one might lose one's job in the next two years, and worries about the security of the current job.[21] Additionally, there is information whether individuals expect an improvement or a worsening of the current job (career) position.[22] It is important to note for the discussion in the following subsection that, except for the earnings' variable, all outcome variables are coded as binary indicators.

The results of the evaluations are given in the following Figures 3 to 7 and Table 4.[23] They show the differences between the control and the ERT group for specific time spans before and after the training for a selected group of outcome variables (multiplied by 100 for outcomes that are indicators).[24] For variables measured using the monthly calendar the distance is expressed in months, for those measured only for the particular month of the yearly interview, the distance is expressed in years.[25] The figures cover up to 18 months or up to 3 'years' before the training and up to 30 months or 3 'years' after ERT. They display the mean effect (solid line; + for the mismatch corrected estimate) and its 95% pointwise confidence interval based on the normal approximation (dashed line; ∇, ∆ for the mismatch corrected estimates). The number of observations available to compute the respective statistics decreases the longer the distance to the incidence of ERT is (see Table 4 for the remaining number of observations). This implies that the variances increase over time. It is reflected in the widening of the confidence intervals. However, the accuracy of the estimated intervals itself may deteriorate, because the normal distribution may not be a very good approximation of the sample distribution when the sample gets too small. Additionally, a

important to account for the monthly labour market pre-ERT history to avoid selecting comparisons observations having too high unemployment probabilities on average.

[21] For non-employed individuals these variables are coded as being "very worried" and as "expecting unemployment".

[22] For non-employed individuals these variables are coded as "expecting no improvement" and "no worsening".

[23] The results presented in the following are based on group 1. The results for group 2 are available on request.

[24] The results for those previously mentioned outcome variables that do not appear here, are not qualitatively different from the ones presented.

mismatch correction may be impossible or very imprecise, because there may be too few observations to identify and estimate the parameters of the ordered probit model.

4.3.2 Results

Figures 3 and 4 present the results of the evaluations for the monthly outcome variables unemployment and full-time employment using comparison group 1.[26] The part left to the 0 vertical mark allows a judgement about the quality of the matches concerning the particular variable.[27] Table 4 presents the accumulated effects for the respective variables.

< --------------------------------------- *Figure 3 about here* -- >

< --------------------------------------- *Figure 4 about here* -- >

Although Figures 3 and 4 as well as Table 4 suggest (almost) significant positive effects of ERT in the short run at least concerning full-time employment, they disappear after about six months. The latter is in line with the findings for public sector sponsored training (Lechner, 1996). It seems to be difficult to reduce the individual unemployment risk by means of training in a rapidly contracting economy that also adjusts to a new economic environment, for example because of unforeseen changes in firm strategies and technologies used, leading to unexpected changes in the size and composition of the work force, so that even previous ERT may only be of limited value.

< ------------------------------------- *Table 4 about here* -- >

< ------------------------------------- *Figure 5 about here* -- >

[25] The time span denoted as the first year is actually the time after the end of ERT and the next interview. Therefore, this time span may vary among individuals. The monthly data is available from July 1989 to December 1993, whereas the yearly data ranges from mid 1990 to early 1994.

[26] *Unemployment* here indicates that the individual has registered for unemployment. There is another monthly variable indicating the receipt of unemployment benefits ("Geld" or "Hilfe"). The results are almost exactly the same when using this second measure of unemployment.

Figure 5 and the right hand part of Table 4 show the two different average effects of ERT regarding monthly gross earnings. From Figure 5 it appears that there are positive effects of about DM 350 from ERT in the second year after completion of the last ERT spell.[28] Note that the same effect appears for the third year, but that probably the reduced sample size leads to its insignificance. The accumulated effects given in Table 4 do not show any significant gain, which is not surprising given the insignificance of an earnings gain in the first year after ERT.

< -- Figure 6 about here -- >

When checking whether the estimated earnings' effects are stable across the population of ERT participants, a significant difference with respect to job position and occupational degree appears. For those individuals having a university degree and / or being in a highly qualified and / or management job position, no significantly positive effects are visible. However for the complements of these populations, positive effects that already appeared in the population average in year two after ERT (Figure 5) are more pronounced (Figures 6 and 7). At least for those not in highly qualified and / or management job positions, there appears a mismatch corrected positive effect for the third year after ERT.[29] Note that these differences of earnings effects regarding education and job position suggest that ERT training and general training are not complementary. Quite the opposite, there appear to be diminishing returns to education.

< -- Figure 7 about here -- >

To see whether the significant earnings gains really translate to additional benefits for participants, information about the participants' share of the direct and indirect training costs is necessary. Results in Lechner (1998) suggest that although the type of training appears not to be too firm specific, most of the cost is paid by the employer. However, it might still be that the

[27] Testing whether these lines deviate significantly from 0 is similar to the tests suggested by Rosenbaum (1984).

[28] The implied average earnings increase is about 9%.

firm is reimbursed by a lower pre- or post-ERT wage. Although exact information on these issues is not available, the previous figures can give some clues. For example, one may wonder whether the insignificant effects in the first year may be due to some sort of sharing the additional productivity between the employer and the employee to make up for training costs, or whether it takes some time for ERT to result in additional productivity. Considering the pre-ERT earnings, it appears that there is no significant cost sharing between employers and employees prior to ERT. Otherwise, the respective figures should show a drop when earnings approach ERT from the left, because post-unification pre-ERT earnings are not part of the conditioning set (attribute vector) for comparison group 1. In conclusion, although there might be some wage restraint in the first ERT year, ERT appears to be very beneficial for earnings. It appears to have no effect concerning the risk of future unemployment. These conclusions are confirmed by considering other outcome variables mentioned in the beginning of this section, in particular those related to future expectations and job position.

The findings with respect to earnings as well as to unemployment are in line with the findings of the already mentioned study by Fitzenberger and Prey (1996). This fact gives some additional confidence in the direction of the effects, because their results are obtained using a different data set and an econometric approach that treats the selection problem very differently. However, compared to the previous studies by Lechner (1996, 1999) investigating off-the-job as well as public-sector-sponsored continuous vocational training, the results differ. Using the same data set and the same econometric methodology, these two papers can not find significant earnings or unemployment effects for the participants of those types of training. This suggests that either the selection of participants into ERT is more efficient than for other types of training, or that ERT itself is a more efficient type of training. Distinguishing between those explanations will be left to future work.

[29] But even for these two groups, the accumulated earnings effects are insignificant, because of the insignificant effect in the first year after ERT.

The sensitivity of the results to different assumptions and specifications is checked in several ways. The results are also computed using the sampling weights provided by the GSOEP, but no qualitative differences appeared. When using *comparison* group 2 - that includes actual job status information that might be considered endogenous - no qualitatively important differences appear. With respect to the earnings variable two issues (unemployment benefits instead of ´0´ for those not working; ln-earnings instead of earnings) are addressed which however confirm the previous results. To check whether there might be differences of the average treatment effects in specific subgroups the sample is divided according to gender, job position, employer characteristics, professional degree, age and pre-training employment status. No important qualitative differences appear, although it is not always possible to determine the earnings' effects significantly, due to the reduced sample sizes. Further checks for sensitivity have been performed with respect to the definition of ERT. No qualitatively important difference appeared. Finally, I considered a *comparison* and treatment group that did not participate in any other form of training, but again there are no important differences to the findings presented here.

5 Conclusion

The general findings of this paper suggest that there are positive earnings effects of ERT participation. However, no corresponding reduction in the average individual unemployment probability of ERT participants can be found. Given that firms appear to cover most of the costs for ERT, it seems very surprising that those individuals face the same risk of being fired as if they would not have participated. This suggests that either their newly acquired firm-specific human capital has lost its value dramatically during the transition process from a centrally planned to a market economy, or that they obtained general skills that increased the worker's value on the labour market, and hence their wage. Hence, they could still be substituted by other workers for more or less the same costs. The latter explanation would be in line with observed gains in post-training earnings. Another explanation compatible with firm-specific training would be that it is very difficult to increase the individual employment probability of workers by means of training in a rapidly contracting economy where mass layoffs are frequent, because changes in the markets are not anticipated by firms. Firms either go bankrupt, thus laying-off the entire work force regardless of their firm-specific human capital, or they close entire production lines with the same loss of all firm-specific human capital to scale back production or to switch to new technologies. Although the latter explanation seems to be more plausible for the specific situation in the GDR, it is beyond the scope of this paper to provide evidence for either of them.

Interesting future research may approach the problem of determining the underlying reasons for the observed effects with samples more suitable for answering these questions. From a technical point of view the differences of the results between the two different *comparison* groups suggest that the selection problems may not yet be solved adequately.

References

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

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

Bell, S.H., L.L. Orr, J.D. Blomquist, and G.G. Cain (1995): *Program Applicants as a Comparison Group in Evaluating Training Programs*, Upjohn: Kalamazoo.

Bera, A., Jarque, C. and Lee, C.F. (1984): "Testing the Normality Assumption in Limited Dependent Variable Models", *International Economic Review*, 25, 563-578.

Blundell, R.W., C. Meghir, and L. Dearden (1995): "The Determinants and Effects of Work Related Training in Britain", *mimeo*, 1995.

Dagenais, M.G. and Dufour, J.M. (1991): "Invariance, Nonlinear Models, and Asymptotic Tests", *Econometrica*, 59, 1601-1615.

Davidson, R. and MacKinnon, J.G. (1984): "Convenient Specification Tests for Logit and Probit Models", *Journal of Econometrics*, 25, 241-262.

Davidson, R., and J.G. MacKinnon (1993): *Estimation and Inference in Econometrics*, Oxford: Oxford University Press.

Dehejia, R., and S. Wahba (1995a): "A Matching Approach for Estimating Causal Effects in Non-Experimental Studies", Harvard University, *mimeo*.

Dehejia, R., and S. Wahba (1995b): "Causal Effects in Non-Experimental Studies", Harvard University, *mimeo*.

Fitzenberger, B., and H. Prey (1996): "Training in East Germany: An Evaluation of the Effects on Employment and Earnings", *unpublished manuscript*, University of Konstanz.

Groot W., J. Hartog, and H. Oosterbeek (1994): "Costs and Revenues of Investments in Enerprise-related Schooling", *Oxford Economic Papers*, 46, 658-675.

Heckman, J.J., and J.A. Smith (1995): "Ashenfelter´s Dip and the Determinants of Participation in a Social Program: Implications for a Simple Program Evaluation Strategies", *Research Report* # 9505, The University of Western Ontario.

Heckman, J.J., and R. Robb (1985):"Alternative Methods of Evaluating the Impact of Interventions", in: J.J: Heckman and B. Singer (eds.), *Longitudinal Analysis of Labour Market Data*, New York: Cambridge University Press.

Heckman, J.J., and V.J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84, 862-880 (includes comments by Holland and Moffitt).

Holland, P.W. (1986): "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-970 (includes comments by Cox, Granger, Glymour, Rubin and a rejoinder by Holland).

LaLonde, R.J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review*, 76, 604-620.

Lechner, M. (1991): "Testing Logit Models in Practice", *Empirical Economics*, 16, 177-198.

Lechner, M. (1996): "An Evaluation of Publicly Funded Continuous Vocational Training in East Germany", *Discussion paper, Beiträge zur angewandten Wirtschaftsforschung # 539-96*, University of Mannheim, revised 1998.

Lechner, M. (1998): *Training the East German Labour Force*, Heidelberg: Physica.

Lechner, M. (1999): "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification", forthcoming in *Journal of Business & Economic Statistics*.

Lynch, L.M. (1992): "Private Sector Training and the Earnings of Young Workers", *The American Economic Review*, 82, 299-312.

Lynch, L.M. (1994): *Training and the Private Sector - International Comparisons*, Chicago: University of Chicago Press.

Newey, W.K., and D.L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing", in Engle, R.F. and D.L. McFadden, Hrsg., *Handbook of Econometrics, Vol. 4,* 2113-2245, Amsterdam: North-Holland.

Orme, C. (1988): "The Calculation of the Information Matrix Test for Binary Data Models", *The Manchaster School*, 56, 370-376.

Orme, C. (1990): "The Small Sample Performance of the Information Matrix Test for Binary Data Models", *Journal of Econometrics*, 46, 309-331.

Rosenbaum, P.R. (1984): "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment", *Journal of the American Statistical Association*, 79, 41-48.

Rosenbaum, P.R. and D.B. Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39, 33-38.

Rosenbaum, P.R., and D.B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrica*, 70, 41-50.

Roy, A.D. (1951): "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 3, 135-146.

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

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

Rubin, D.B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

Sobel, M.E. (1994): "Causal Inference in the Social and Behavioral Sciences", in: G. Arminger, C.C. Clogg and M.E. Sobel (eds.): *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, New York: Plenum Press.

Wagner, G.G., R.V. Burkhauser, and F. Behringer (1993): "The English Language Public Use File of the German Socio Economic Panel", *Journal of Human Resources*, 28, 429-433.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1-25; "Corrigendum", *Econometrica*, 51, 513.

Appendix A: Econometrics

A.1 Matching protocol

This section gives the details of the matching protocol used for the final evaluations as proposed by Lechner (1999). For further discussion of its properties the reader is refered to that paper.

Step 1: Split observations in two exclusive pools according to whether they participated in ERT (T-pool) or not (C-pool).

Step 2: Draw randomly an observation in T-pool (denoted by *i*) and remove from T-pool.

Step 3: Define calliper of partial propensity score for observation *i* in terms of the predicted index $v_i\hat{\beta}$ and its conditional variance $Var(V\hat{\beta} \mid V = v_i)$. The latter is derived from $\mathrm{Var}(\hat{\beta})$ by the delta method.

Step 4: Find observations in C-pool (denoted by *j*) obeying $v_j\hat{\beta} \in [v_i\hat{\beta} \pm c\sqrt{Var(v_i\hat{\beta})}]$. The constant *c* is chosen such that the interval is identical to a 50% confidence interval around $v_i\hat{\beta}$.

Step 5: (a) If there is only one or no observation in this interval: find observation *j* in C-pool that is closest to observation *n*, such that it minimizes $(v_j\hat{\beta} - v_i\hat{\beta})^2$.

(b) If there are two or more observations in this set generated by Step 4: Take these controls and compute the variables *m* in relation to the starting date of observation *n*. Denote these and perhaps other variables that are already included in *V* as $\tilde{m}_j$ and $\tilde{m}_i$, respectively. Define a distance between each control *j* and *n* as $d(j,i) = (v_j\hat{\beta}, \tilde{m}_j)' - (v_i\hat{\beta}, \tilde{m}_i)'$. Choose control *j* such that it has the smallest Mahalanobis distance $m(j,i) = d(j,i)'Wd(j,i)$ within the calliper. *W* denotes the inverse of the estimated variance of $(v\hat{\beta}, \tilde{m})'$ in the C-pool.

Step 6: Remove *j* from C-pool.

Step 7: If there are any observations in the T-pool left, start again with step 2.

A.2 Correction for mismatches

This appendix briefly gives the method used to correct for any mismatch remaining after using the algorithm described in section A.1 of this appendix (for more details see Lechner, 1999). Denote the difference of the potential outcomes by $\Delta Y\ = (Y^t - Y^n)$. The realizations of the sample and the matching process gives us pairs ($\Delta y_i, \Delta x_i$), $i \le N^t$. Define the difference in terms of balancing scores as $\Delta b(x_i) = b(x_i) - b(x_j)$. x_j denotes the value of x for observation *j* that is matched to observation *i*. Note that in general, equation (A.1) holds:

$$E[\Delta y_i \,|\, \Delta b(x_i) = 0] = E\{E[(Y^t - Y^n) \,|\, b(X) = b(x_i)] \,|\, S = 1\} = E\{E[\Delta Y \,|\, b(X) = b(x_i)] \,|\, S = 1\} = \theta^0. \quad \text{(A.1)}$$

However, in a finite sample Δx_i may not be exactly zero. In the case of continuous variables it seems reasonable to assume that the conditional expectation of the dependent variable is linear in $\Delta b(x_i)$, because matching already removed most differences in $\Delta b(x_i)$:

$$E[\Delta y_i \,|\, \Delta b(x_i) = \eta_i] = \theta^0 + \eta_i \lambda, \quad i \le N^t. \qquad \text{(A.2)}$$

θ^0 can be estimated by regressing Δy_i on $\Delta b(x_i)$ and a constant (cf. Rubin, 1979).[30]

Suppose now that the outcome consists of only two values (0, 1), hence the support of ΔY is the set {-1,0,1}. In this case, the treatment effect can be written as:

$$\theta^0 = E(\Delta Y | S = 1) = P(\Delta Y = 1 | S = 1) - P(\Delta Y = -1 | S = 1) \qquad \text{(A.3)}$$

A consistent estimate of the average treatment effect can be obtained by substituting sample analogues for the population probabilities:

[30] Standard errors are computed using a heteroscedasticity robust estimator. The particular variant is labelled as HC_2 by Davidson and MacKinnon (1993, p.554) and has good small sample properties.

$$\hat{\theta}_{N^t} = \frac{1}{N^t}\sum_{i=1}^{N^t}[P(\Delta y_i = 1 \mid \Delta b(x_i) = 0) - P(\Delta y_i = -1 \mid \Delta b(x_i) = 0)] \qquad \text{(A.4)}$$

In a first step a three-group-ordered probit model is estimated with Δy_i as dependent variable and $\Delta b(x_i)$ plus a constant as independent variables. In the second step, the above probabilities are directly derived from this model and computed for the individual observations using the estimated coefficients. Finally, the variance of $\hat{\theta}_{N^t}$ is approximated from the variance of the estimated coefficients by the delta method.

Appendix B: Specification tests for the partial propensity score

Conditional homoscedasticity (implied by independence) and normality are tested using conventional specification tests (score-tests similar to Bera, Jarque, and Lee, 1984, Davidson and MacKinnon, 1984, and Orme, 1988, 1990).[31] The alternative hypothesis for the heteroscedasticity test is that the variance of the error term varies with a particular variable.[32] Furthermore, the consistency property of the specification tests, in particular of such omnibus tests as the information matrix test (IMT) will eventually detect any other dependence of $V\beta^0$ and *f(M,U)*. Several versions of the IMT are computed: the full version using *all indicators* available. This is the omnibus test, which however sometimes has a tendency of excess rejection in small samples. Therefore, the *only main diagonal indicators*-version of the IMT uses only elements from the main diagonal of the difference between OPG and minus expected hessian (jointly). For problems related to a particular variable (such as random coefficient variation) IMT statistics using only a single main diagonal indicator are powerful

[31] The use of semiparametric methods has been considered. However, it is not necessary, because the specification tests indicate no violation of the distributional assumptions necessary for the probit model.

[32] $U = U^* \exp(V^j\alpha^0), U^* \mid V = v \sim N(0,1); \quad H^0 : \alpha^0 = 0; \quad H^1 : \alpha^0 \neq 0.$ Note that the alternative hypothesis is observationally equivalent to a probit model with all coefficients varying with $[1/ exp(V^j\alpha^0)]$. V^j denotes a particular element of the vector *V*, and α^0 denotes the true value of an unknown coefficient vector.

tests (Lechner, 1991). For the details of the computations the reader is referred to the literature mentioned above.

The t-values and score test results against heteroscedasticity presented in Table B.1 are computed using the GMM (or PML) formula given in White (1982).[33] The information-matrix-tests statistics are computed using the second version suggested in Orme (1988).[34]

[---------------------- include Table B.1 and B.2 about here ------------------]

[33] Five versions are computed: based on the matrix of the outer product of the gradient (OPG) alone, on the empirical hessian alone, on the expected (under the null) hessian alone, and on combining the hessian, respectively the expected hessian (under the null), and the OPG. Previous Monte Carlo studies (e.g. Davidson and MacKinnon, 1984, Lechner, 1991) as well as theoretical papers (e.g. Dagenais and Dufour, 1991) show that tests based on the latter at least avoid some undesirable properties which can occur with other versions. Therefore, the results presented are computed using these estimates of the covariance matrix.

[34] The first version is almost numerically identical. *Only main-diagonal indicators* refers to a version of the information matrix test using as test indicators only the main diagonal of the difference between OPG matrix and the matrix of the expected hessian.