*J. R. Statist. Soc.* A (2002)
**165**, *Part* 1, *pp.* 59–82

# Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods

Michael Lechner

*University of St Gallen, Switzerland*

[Received July 2000. Final revision September 2001]

**Summary.** Recently several studies have analysed active labour market policies by using a recently proposed matching estimator for multiple programmes. Since there is only very limited practical experience with this estimator, this paper checks its sensitivity with respect to issues that are of practical importance in this kind of evaluation study. The estimator turns out to be fairly robust to several features that concern its implementation. Furthermore, the paper demonstrates that the matching approach *per se* is no panacea for solving all the problems of evaluation studies, but that its success depends critically on the information that is available in the data. Finally, a comparison with a bootstrap distribution provides some justification for using a simplified approximation of the distribution of the estimator that ignores its sequential nature.

*Keywords*: Balancing score; Matching; Multiple programmes; Programme evaluation; Sensitivity analysis; Treatment effects

## 1. Introduction

Many European countries use substantial active labour market policies (ALMPs) to bring Europe’s notoriously high levels of unemployment back to some sort of socially acceptable level by increasing the employability of the unemployed. These policies consist typically of a variety of subprogrammes, such as employment programmes, training and wage subsidies, among others.

Recent evaluation studies surveyed for example by Fay (1996) and Heckman *et al.* (1999) do not appear to develop any consensus on whether these programmes are effective for their participants. On the contrary, many studies raise serious doubts. However, it could be argued that the policy implications of many of these studies were limited because their econometric framework was not ideally suited to the problem, and because the available data that were used were typically far from being ideal as well.

Recently the Swiss Government encouraged several groups of researchers to evaluate the Swiss ALMPs by using administrative data from the unemployment registers and the pension system. Among those studies were also two econometric studies by Lalive *et al.* (2000) and Gerfin and Lechner (2000). The first used a structural econometric modelling approach based on modelling the duration of unemployment, whereas the second used an extension of an essentially nonparametric pseudoexperimental matching approach to multiple treatments proposed and discussed by Lechner (2001a,b). The fact that these studies used different (more

*Addresses for correspondence*: Michael Lechner, Swiss Institute for International Economics and Applied Economic Research, University of St Gallen, Dufourstrasse 48, CH-9000 St Gallen, Switzerland.
E-mail: Michael.Lechner@unisg.ch

0964–1998/02/165059

or less explicit) identification strategies points to the issue that for every evaluation study there is the crucial question of which identification strategies and estimation method would be suitable for the specific situation. Angrist and Krueger (1999), Heckman and Robb (1986) and Heckman *et al.* (1999) provide an excellent overview of the available identification and resulting estimation strategies.

Of course the choice of an identification strategy is strongly linked to the type of data that are available about the selection process for the programmes. Gerfin and Lechner (2000) argue that they observe the major variables influencing selection as well as outcomes, so the assumption that labour market outcomes and selection are independent conditionally on these observables (the conditional independence assumption (CIA)) is plausible. Being able to use a CIA for identification in combination with having a large data set has implications for the choice of a suitable estimator. The desirable properties of an estimator in this situation are that it should avoid almost any other assumption than the CIA, such as functional form assumptions for specific conditional expectations of the variables of interest. In particular the estimator of choice should avoid restricting the effects of the programmes to be the same in specific subpopulations because there is substantial *a priori* evidence that those programmes could have very different effects for different individuals (effect heterogeneity). Finally, this ideal estimator must take account of the very different programmes that make up the Swiss ALMPs (programme heterogeneity). To be able to convince policy makers about the merits of the results of any evaluation, the estimator needs to be based on a general concept that could easily be communicated to non-econometricians.

An estimator that is nonparametric in nature allows for effect as well as programme heterogeneity, and one that is based on a statistical concept that is easy to communicate is the recently suggested matching estimator for heterogeneous programmes. The general idea of matching is to construct an artificial comparison group. The average labour market outcomes of this group are compared with the average labour market outcomes of the group of participants in the programme. When the CIA is valid, this estimator is consistent when the selected comparison group and the group in the specific programme have the same distribution of observable factors determining jointly labour market outcomes and participation. Matching for binary comparisons has recently been discussed in the literature and applied to various evaluation problems by Angrist (1998), Dehejia and Wahba (1999), Heckman *et al.* (1997, 1998), Lechner (1999, 2000) and Smith and Todd (2000), among others. The standard matching approach that considers only two states (for example in the programme compared with not in the programme) has been extended by Imbens (2000) and Lechner (2001b) to allow for multiple programmes.

The results by Gerfin and Lechner (2000) indicate considerable heterogeneity with respect to the effects of different programmes. They find substantial positive employment effects for one particular programme that is a unique feature of the Swiss ALMPs. It consists of a wage subsidy for temporary jobs in the regular labour market that would otherwise not be taken up by the unemployed. They also find large negative effects for traditional employment programmes operated in sheltered labour markets. For training courses the results are mixed.

There is only very limited practical experience with these kinds of matching estimator for multiple programmes (to the best of our knowledge, the only other applications of this specific approach are Brodaty *et al.* (2001), Dorsett (2001), Frölich *et al.* (2000), Larsson (2000) and Lechner (2001a). In particular Lechner (2001a) discusses issues that are relevant for the implementation of the estimator. Here we cover several other points that could be potentially responsible for the results that were obtained by Gerfin and Lechner (2000). It is of particular interest whether the stark differences between the effects for the two different types of subsidized employment are robust in these respects. In addition, the sensitivity of the results to the amount of information that is included in the estimation will be addressed. Obviously, robustness of the results should not be expected in that exercise.

The plan of this paper is as follows. The next section summarizes the results for multiple treatments that were obtained in Lechner (2001b) and describes the estimator proposed. Section 3 briefly discusses several aspects of the application. Section 4 presents the results of the base-line specification. Section 5 discusses the sensitivity of the results by considering several deviations from the base-line specification. Section 6 concludes.

## 2. Econometric framework for the estimation of the causal effects

### 2.1. Notation and definition of causal effects

#### 2.1.1. Notation

The prototypical model of the microeconometric evaluation literature is the following. An individual can choose between two states (causes). The potential participant in a programme receives a hypothetical outcome (e.g. earnings) in both states. This model is known as the Roy (1951) and Rubin (1974) model of potential outcomes and causal effects (see Holland (1986) for an extensive discussion of concepts of causality in statistics, econometrics and other fields).

Consider the outcomes of $M+1$ different mutually exclusive states denoted by $\{Y^0, Y^1, \ldots, Y^M\}$. Following that literature the different *states* are called *treatments*. It is assumed that each individual receives only one of the treatments. Therefore, for any individual, only one component of $\{Y^0, Y^1, \ldots, Y^M\}$ can be observed in the data. The remaining $M$ outcomes are counterfactuals. Participation in a particular treatment $m$ is indicated by the variable $S \in \{0, 1, \ldots, M\}$.

#### 2.1.2. Pairwise effects

Assuming that the typical assumptions of the Rubin model are fulfilled (see Holland (1986) or Rubin (1974), for example), equation (1) defines pairwise average treatment effects of treatments $m$ and $l$ for the participants in treatment $m$:

$$\theta_0^{m,l} = E(Y^m - Y^l|S=m) = E(Y^m|S=m) - E(Y^l|S=m). \tag{1}$$

$\theta_0^{m,l}$ denotes the expected effect for an individual randomly drawn from the population of participants in treatment $m$. If participants in treatments $m$ and $l$ differ in a way that is related to the distribution of attributes (or exogenous confounding variables) $X$, and if the treatment effects vary with $X$, then $\theta_0^{m,l} \neq -\theta_0^{l,m}$, i.e. the treatment effects on the treated are not symmetric.

### 2.2. Identification

#### 2.2.1. The conditional independence assumption

The framework set up above clarifies that the average causal effect is generally not identified. Therefore, this lack of identification must be overcome by plausible untestable assumptions. Their plausibility depends on the problem that is being analysed and the data that are available. Angrist and Krueger (1999), Heckman and Robb (1986) and Heckman *et al.* (1999) provide an excellent overview about identification strategies that are available in different situations.

Imbens (2000) and Lechner (2001b) considered identification under the CIA in the model with multiple treatments. A CIA defined to be valid in a subspace of the attribute space is formalized by

$$Y^0, Y^1, \ldots, Y^M \coprod S|X = x, \qquad \forall x \in \chi. \tag{2}$$

This assumption requires the researcher to observe all characteristics that jointly influence the outcomes as well as the selection for the treatments. In that sense, the CIA may be called a 'data hungry' identification strategy. Note that the CIA is not the minimal identifying assumption, because all that is needed to identify mean effects is conditional mean independence. However, the CIA has the virtue of making the latter valid for all transformations of the outcome variables. Furthermore, in most empirical studies it would be difficult to argue why conditional mean independence should hold and CIA might nevertheless be violated.

In addition to independence it is required that all individuals in that subspace actually can participate in all states (i.e. $0 < P(S = m|X = x), \forall m = 0, \ldots, M, \forall x \in \chi$). This condition is called the common support condition and is extensively discussed in Lechner (2001c). For any pairwise comparison it is sufficient that, for all values of $X$ for which those treated have positive marginal probability, there could be comparison observations as well.

Lechner (2001b) shows that the CIA identifies the effects defined in equation (1). Indeed, Gerfin and Lechner (2000) argued that their data are so rich that it seems plausible to assume that all important factors that jointly influence labour market outcomes and the process selecting people for the different states can be observed. Therefore, the CIA is the identifying assumption of choice. In Section 4 we elaborate on the actual identification in this application.

### 2.2.2. *Reducing the dimension by using balancing scores*

In principle the basic ingredients of the final estimator would be estimators of expressions like $E(Y^l|X, S = l)$, because the CIA implies that $E(Y^l|S = m) = E_X\{E(Y^l|X, S = l)|S = m\}$. However, nonparametric estimators could be problematic, because of the potentially high dimensional $X$ and the resulting so-called *curse of dimensionality*. For two treatments, however, Rosenbaum and Rubin (1983) showed that conditioning the outcome variable on $X$ is not necessary, but it is sufficient to condition on a scalar function of $X$, namely the participation probability conditional on the attributes (this is the so-called balancing score property of the propensity score). For the case of multiple treatments Lechner (2001b) shows that some modified versions of the balancing score properties hold in this more general setting as well.

Denote the marginal probability of treatment $j$ conditional on $X$ as $P(S = j|X = x) = P^j(x)$. Lechner (2001a) shows that the following result holds for the effect of treatment $m$ compared with treatment $l$ on the participants in treatment $m$:

$$\begin{aligned} \theta_0^{m,l} &= E(Y^m|S = m) - \underset{P^{l|ml}(X)}{E}[E\{Y^l|P^{l|ml}(X), S = l\}|S = m]. \\ P^{l|ml}(x) &= P^{l|ml}(S = l|S \in \{l, m\}, X = x) = \frac{P^l(x)}{P^l(x) + P^m(x)}. \end{aligned} \tag{3}$$

If the respective probabilities $P^{l|ml}(x)$ are known or if a consistent estimator is available, the dimension of the estimation problem is reduced to 1. If $P^{l|ml}(x)$ is modelled directly, no information from subsamples other than those containing participants in $m$ and $l$ is needed for the identification and estimation of $\theta_0^{m,l}$ and $\theta_0^{l,m}$. Thus, we are basically back in the binary treatment framework.

In many evaluation studies considering multiple exclusive programmes it is natural to specify jointly the choice of a particular treatment from all or a subset of available options. $P^{l|ml}(x)$ could then be computed from that model. In this case, consistent estimates of all marginal choice probabilities $[P^0(X), \ldots, P^M(X)]$ can be obtained. Hence, it may be attractive to condition jointly on $P^l(X)$ and $P^m(X)$ instead of on $P^{l|ml}(X)$. $\theta_0^{m,l}$ is identified in this case as well, because $P^l(X)$ together with $P^m(X)$ are 'finer' than $P^{l|ml}(X)$:

$$E\{P^{l|ml}(X)|P^l(X), P^m(X)\} = E\left\{\frac{P^l(X)}{P^l(X) + P^m(X)}|P^l(X), P^m(X)\right\} = P^{l|ml}(X). \tag{4}$$

### 2.3. A matching estimator

Given the choice probabilities, or a consistent estimate of them, the terms appearing in equations (3) can be estimated by any parametric, semiparametric or nonparametric regression method. One of the popular choices of estimators in a binary framework is matching (for recent examples see Angrist (1998), Dehejia and Wahba (1999), Heckman *et al.* (1998), Lechner (1999, 2000) and Smith and Todd (2000)). The idea of matching on balancing scores is to estimate $E(Y^l|S = m)$ by forming a comparison group of selected participants in $l$ that has the same distribution for the balancing score (here $P^{l|ml}(X)$ or $[P^l(X), P^m(X)]$) as the group of participants in $m$. By virtue of the property of being a balancing score, the distribution of $X$ will also be balanced in the two samples. The estimator of $E(Y^l|S = m)$ is the mean outcome in that selected comparison group. Typically, the variances are computed as the sum of empirical variances in the two groups (ignoring the way that the groups have been formed). Compared with nonparametric regression estimates, a major advantage of matching is its simplicity and its intuitive appeal. The advantages compared with parametric approaches are its robustness to the functional form of the conditional expectations (with respect to $E(Y^l|X, S = l)$) and that it leaves the individual causal effect completely unrestricted and hence allows arbitrary heterogeneity of the effects in the population. Lechner (2001a,b) proposes and compares different matching estimators that are analogous to the rather simple matching algorithms used in the literature on binary treatments. The exact matching protocol that is used for the application is based on $[P^l(X), P^m(X)]$ and is detailed in Table 1.

Several comments are necessary. Step 2 ensures that we estimate only effects in regions of the attribute space where two observations from two treatments can be observed having a similar participation probability (the common support requirement). Otherwise the estimator will give biased results (see Heckman *et al.* (1998)).

A second remark with respect to the matching algorithm concerns the use of the same comparison observation repeatedly in forming the comparison group (*matching with replacement*). This modification of the 'standard' estimator is necessary for this estimator to be applicable at all when the number of participants in treatment $m$ is larger than in the comparison treatment $l$. Since the role of $m$ and $l$ could be reversed in this framework, this will always be the case when the number of participants is not equal in all treatments. This procedure has the potential problem that a few observations may be heavily used although other very similar observations are available. This may result in a substantial and unnecessary inflation of the variance. Therefore, the potential occurrence of this problem should be monitored.

**Table 1.** Matching protocol for the estimation of $\theta_0^{m,l}$

| *Step* | *Description* |
|---|---|
| 1 | Specify and estimate a multinomial probit model to obtain $[\hat{P}_N^0(x), \hat{P}_N^1(x), \dots, \hat{P}_N^M(x)]$ |
| 2 | Restrict sample to common support: delete all observations with probabilities larger than the smallest maximum and smaller than the largest minimum of all subsamples defined by $S$ |
| 3 | Estimate the respective (counterfactual) expectations of the outcome variables. For a given value of $m$ and $l$ the following steps are performed:<br>(a) choose one observation in the subsample defined by participation in $m$ and delete it from that pool;<br>(b) find an observation in the subsample of participants in $l$ that is as close as possible to that chosen in step (a) in terms of $[\hat{P}_N^m(x), \hat{P}_N^l(x), \tilde{x}]$; $\tilde{x}$ contains information on sex, duration of unemployment, native language and start of programme; 'closeness' is based on the Mahalanobis distance; do not remove that observation, so that it can be used again;<br>(c) repeat (a) and (b) until no participant in $m$ is left;<br>(d) using the matched comparison group formed in (c), compute the respective conditional expectation by the sample mean; note that the same observations may appear more than once in that group |
| 4 | Repeat step 3 for all combinations of $m$ and $l$ |
| 5 | Compute the estimate of the treatment effects using the results of step 4 |

A third remark concerns the appearance of the variables $\tilde{x}$ in step 3(b). This subset of conditioning variables already appears in the score. The motivation for also including them explicitly in the matching is that they are potentially highly correlated with the outcome variables (but not influenced by them) as well as with selection. Therefore, it seems to be particularly important to obtain very good matches with respect to these variables even in smaller samples. However, by virtue of the balancing score property, including them as additional matching variables is not necessary asymptotically because they are already included in the score. Note that including them in the score as well as additional matching variables amounts to increasing the weight of these variables, which is suspected to be critically important, when forming the matches.

## 3. Application

The application in this paper is based on the evaluation study of the various programmes of the Swiss ALMPs by Gerfin and Lechner (2000). They focused on the individual success in the labour market that is due to these programmes. The Swiss Government made available a very informative and large database consisting of administrative records from the unemployment insurance system as well as from the social security system. It covers the population of unemployed people in December 1997. Gerfin and Lechner (2000) claim that in these data all major factors that jointly influence both the selection for the various programmes as well as employment outcomes are observed.

Let us very briefly reconsider their main line of argument to establish identification. First note that the decision to participate in a programme is made by the case-worker according to his impressions obtained mainly from the monthly interviews of the unemployed. To evaluate this 'subjective impression' the law requires that programmes must be necessary and adequate to improve individual employment chances. Although the final decision about participation is always made by the case-worker (or somebody whom the case-worker must report to), the unemployed may also try to influence this decision during the conversations that take place in these interviews. Furthermore, although the law is enacted at the federal level, the 26 Swiss cantons exercise considerable autonomy in interpreting and implementing the rules that are specified in this law. To summarize, it does not appear to be possible to state exactly how an individual participation decision is made, but it should be possible to specify the information set on which this decision is based. Luckily, all the information that is obtained by and available to the case-worker is stored in a centralized database to which we have access and which is described below. To that data coming from the unemployment registrars we add information on the last 10 years of labour market history coming from the pension system. We suspect that labour market experience influences the individual preferences considerably, although it might be argued that the relevant part for selection and outcome is already contained in the database coming from the unemployment registrar. In the following the database and the sample, as well as the programmes, are briefly described.

The data from the unemployment registrars cover the period from January 1996 to March 1999 for all individuals who were registered as unemployed on December 31st, 1997. These data provide very detailed information about the unemployment history, ALMP participation and personal characteristics. The pension system data cover 1988–1997 for a random subsample of about 25000 observations. The exact variables used in this study can be found in appendix WWW that can be downloaded from the Internet:

```
http://www.siaw.unisg.ch/lechner/l_jrss_a
```

They cover sociodemographics (age, gender, marital status, native language, nationality, type of work permit and language skills), region (town or village and labour office), subjective valuations by the case-worker (qualifications and chances of finding a job), sanctions imposed by the placement office, previous jobs and job desired (occupation, sector, position, earnings and full or part time), a short history of labour market status on a daily basis, and the employment status and earnings on a monthly basis for the last 10 years. Gerfin and Lechner (2000) applied a series of sample selection rules to the data. The most important are to consider only individuals who were unemployed on December 31st, 1997, with a spell of unemployment of less than 1 year who have not participated in any major programme in 1997 and are aged between 25 and 55 years.

The ALMPs can be grouped into three broad categories:

(a) training courses,
(b) employment programmes EP and
(c) temporary employment with wage subsidy TEMP.

The first two groups are fairly standard for a European ALMP encompassing a variety of programmes. The last type of programme is quite unique, however. The difference between (b) and (c) is that employment programmes take place outside the 'regular' labour market (see below). By contrast TEMP refers to a regular job.

In this study we focus on a subset of programmes, namely computer courses COC, EP and TEMP (and non-participation NONP) (the first participation in a programme with a duration of more than 2 weeks, starting after January 1st, 1998, decides the assignment to the appropriate group; any participation in a programme later is treated as being the effect of the first programme). The effects for these programmes were the most interesting ones found in Gerfin and Lechner (2000). Note that the validity of the CIA allows us to analyse the effects of these programmes on the subsample of non-participants and participants in the respective programmes, thus avoiding any selectivity bias problems that arise from ignoring individuals in other programmes that are not considered here. The reduction of the sample has the important advantage for this paper that computation times are considerably reduced.

A problem concerns the group of non-participants. For this group important time-varying variables like 'duration of unemployment before the programme' are not defined. To make meaningful comparisons with those unemployed people entering a programme, in the base-line estimate an approach suggested in Lechner (1999) is used: for each non-participant a hypothetical programme starting date from the sample distribution of starting dates is drawn. People with a simulated starting date that is later than their actual exit date from unemployment are excluded from the data set. Later in Section 5.1 other ways to handle this problem will be presented. Note that deleting non-participants could potentially bias the results of the effects of the programmes on non-participants, because it changes the distribution of non-participants by deleting systematically the data for individuals with higher unemployment probabilities. However, this has no implication for effects defined for any of the populations of participants, which are typically those of interest with regard to policy.

Table 2 shows the number of observations as well as some descriptive statistics for subsamples composed of non-participants as well as participants in the three programme groups that were considered. The mean duration of the programme is just 1 month for computer courses and almost 150 days for employment programmes. Table 2 shows that important variables like qualifications, nationality and duration of unemployment also vary substantially. The final column indicates that the employment rate at the last day in our data varies considerably between 26% and 48%. Of course, this is not indicative of the success of a programme because the composition of different groups of participants differs substantially with respect to variables influencing future employment, so we expect differences for these different groups of unemployed even when they would not have participated in any programme.

## 4. Results for the base-line scenario

### 4.1. Selection for the programmes

The base-line scenario basically reproduces the results that were obtained by Gerfin and Lechner (2000) for the sample used here. The first step is an estimation of the conditional probabilities of ending in each of the four states. The full set of the estimation results of a multinomial probit model using simulated maximum likelihood with the Geweke–Hajivassiliou–Keane (GHK) simulator and 200 draws for each observation and choice equation (e.g. Börsch-Supan and Hajivassiliou (1993) and Geweke *et al.* (1994)) can be found in appendix WWW:

http://www.siaw.unisg.ch/lechner/l_jrss_a

**Table 2.** Number of observations and selected characteristics of different groups†

| *Group* | *Observations (persons)* | *Duration of programme (mean days)* | *Unemployment before (mean days)* | *Qualification (mean)* | *Foreign (share, %)* | *Employed March 1999 (share, %)* |
|---|---|---|---|---|---|---|
| NONP | 6735 | 0 | 250‡ | 1.8 | 47 | 39 |
| COC | 1394 | 36 | 214 | 1.3 | 22 | 44 |
| EP | 2473 | 147 | 300 | 1.8 | 46 | 26 |
| TEMP | 4390 | 114 | 228 | 1.7 | 46 | 48 |

†Qualification is measured as 1, skilled, 2, semiskilled, and 3, unskilled.
‡Start date simulated.

The variables that are used in the multinomial probit model are selected by a preliminary specification search based on binary probits (each relative to the reference category NONP) and score tests against variables omitted. The final specification contains a varying number of mainly discrete variables that cover groups of attributes related to personal characteristics, valuations of individual skills and chances in the labour market as assessed by the placement office, previous and desired future occupations, and information related to the current and previous spell of unemployment, and past employment and earnings. Variables that are only related to selection and not to the potential outcomes need not be included for consistent estimation.

In practice, some restrictions on the covariance matrix of the errors terms of the multinomial probit model need to be imposed, because not all elements of it are identified and to avoid excessive numerical instability. Here all correlations of the error terms with the error term of the reference category are restricted to zero. The covariance matrix is not estimated directly, but the corresponding Cholesky factors are used.

The results are very similar to those obtained by Gerfin and Lechner (2000), to which the reader is hence referred for the detailed interpretation. Here it is sufficient to note that there is considerable heterogeneity with respect to the selection probabilities. Again we find that better 'risks' (in terms of unemployment risk) are more likely to be in COC, whereas 'bad risks' are more likely to be observed in EP.

Table 3 shows descriptive statistics of the estimated probabilities that are the basis for matching. In particular there is a large negative correlation between the probabilities of TEMP and EP with NONP.

### 4.2. Matching

The numbers of observations deleted because of the common support requirement across different subsamples are given in Table 4. The criterion that is used is that all estimated marginal probabilities are larger than the smallest maximum of the corresponding probability in any sample. The reverse must hold for minima. The share of observations that are lost varies between subsamples, but they are very small, never exceeding 3% in this paper. In contrast, Gerfin and Lechner (2000) found a reduction of more than 14% due to so-called language courses whose participants are very different from the rest of the unemployed. These courses have been omitted from the current analysis. For a detailed discussion of issues related to the common support problem, see Lechner (2001c).

Since one-to-one matching is with replacement, there is the possibility that an observation may be used many times, thus inflating the variance. Table 5 presents the share of the weights

**Table 3.** Descriptive statistics of the predicted probabilities from the multinomial probit model

| *Group* | *Mean (%)* | *Standard deviation × 100* | *Correlations* | | | |
|---|---|---|---|---|---|---|
| | | | *NONP* | *COC* | *EP* | *TEMP* |
| NONP | 44.9 | 12.98 | 1 | −0.21 | −0.48 | −0.52 |
| COC | 9.3 | 8.55 | | 1 | −0.32 | −0.19 |
| EP | 16.4 | 11.47 | | | 1 | −0.22 |
| TEMP | 29.3 | 10.88 | | | | 1 |

**Table 4.** Loss of observations due to the common support requirement†

| *Group* | *Observations before* | *Observations after* | *% deleted* |
|---|---|---|---|
| NONP | 6735 | 6575 | 3 |
| COC | 1394 | 1375 | 1 |
| EP | 2473 | 2419 | 2 |
| TEMP | 4390 | 4258 | 3 |

†The total number of observations decreases by 365 owing to the enforcement of the common support requirement.

**Table 5.** Share of the largest 10% of the weights to total weight (number of participants)†

| *Group* | *Shares (%) for the following groups:* | | | |
|---|---|---|---|---|
| | *NONP* | *COC* | *EP* | *TEMP* |
| NONP | | 41 | 35 | 27 |
| COC | 21 | | 33 | 24 |
| EP | 24 | 42 | | 24 |
| TEMP | 24 | 42 | 35 | |

†Observations from the sample denoted in the column are matched to observations of the sample denoted in the row.

of the 10% of observations that have been used most (i.e. 10% of those matched comparisons with the largest weights are matched to *number-in-table percentage* of the treated; this concentration ratio must of course be larger than 10% which corresponds to the case when every comparison observation is used only once). Given the limited experience with this approach the respective numbers appear to be in the usual range. It is obvious, however, that the smaller the sample the smaller the diversity of the probabilities so the same observations are used more frequently.

Checking the quality of match with respect to several variables including the probabilities used for matching shows that the matched comparison samples are very similar to the treated samples.

### 4.3. Effects

The measure of the success of the programme is employment in the regular labour market at any given time after the start of the programme. Hence the outcome variable is binary. The time on the programme is not considered to be regular employment. Owing to the limitations of the data the potential period of observing programme effects cannot be longer than 15 months, because the latest observation dates from March 31st, 1999. In that sense the analysis will be restricted to the short run effects of the ALMP.

Table 6 displays the mean effects of the programmes on their respective participants 1 year after the individual participation in the programme starts. The entries on the main diagonal

**Table 6.** Average effects for participants ($\theta_0^{m,l}$) measured as the difference in employment rates 1 year after the start of the programme†

| *Group m* | *Differences in employment rates (percentage points) for the following groups l:* | | | |
|---|---|---|---|---|
| | *NONP* | *COC* | *EP* | *TEMP* |
| NONP | 40.7 | 2.1 (3.2) | **7.2** (2.3) | **−6.4** (1.6) |
| COC | **−8.3** (2.5) | 45.9 | −2.1 (3.5) | **−9.1** (2.7) |
| EP | **−8.4** (2.3) | −6.5 (4.1) | 30.9 | **−15.7** (2.5) |
| TEMP | *4.2* (1.7) | **8.0** (3.3) | **13.8** (2.7) | 50.1 |

†Standard errors are given in parentheses. Results are based on matched samples. Numbers in bold indicate significance at the 1% level (two-sided test); numbers in italics indicate significance at the 5% level. Unadjusted levels lie on the main diagonal.

show the employment rates in the four groups in percentage points. The programme effects are off the main diagonals (for simplicity in most cases NONP is called a programme). A positive number indicates that the effect of the programme shown in the row compared with the programme appearing in the column is an on-average higher rate of employment for those who participate in the programme given in the row (for example, the mean effect of TEMP compared with COC is 8.0 percentage points of additional employment for participants in TEMP).

The results for the respective participants in the programmes (the upper part of Table 5) indicate that TEMP is superior to almost all the other programmes. The mean gain compared with the other programmes is between about 6 and 16 percentage points. In particular TEMP is the only programme that dominates NONP. In contrast, EP has negative effects. COC is somewhat intermediate in general, but the COC programmes do look fairly bad for their participants.

Fig. 1 shows the dynamics of the effects by pinning down their development over time after the start of the programme. It presents the pairwise effects for all programmes and their respective participants. A value larger than 0 indicates that participation in the programme would increase the chances of employment compared with being allocated to the other programme in question.

Considering the relative positions of the curves, the line for NONP reveals the expected profile (Figs 1(a)–1(c)): in the beginning it is positive and increasing, but then it starts to decline as participants leave their respective programmes and increase their job search activities. Overall the findings set out in Table 6 are confirmed: TEMP dominates. EP is dominated by NONP and TEMP. For those participating in EP there is no significant difference compared with participating in COC. For the participants in COC there is a small positive initial effect compared with EP. This effect is probably because COC programmes are much shorter than EP programmes.

## 5. Sensitivity analysis

There is only very limited practical experience with these kinds of matching estimator for multiple programmes. In particular Lechner (2001a) discusses several topics that are relevant for the implementation of the estimator. Here, these considerations are extended

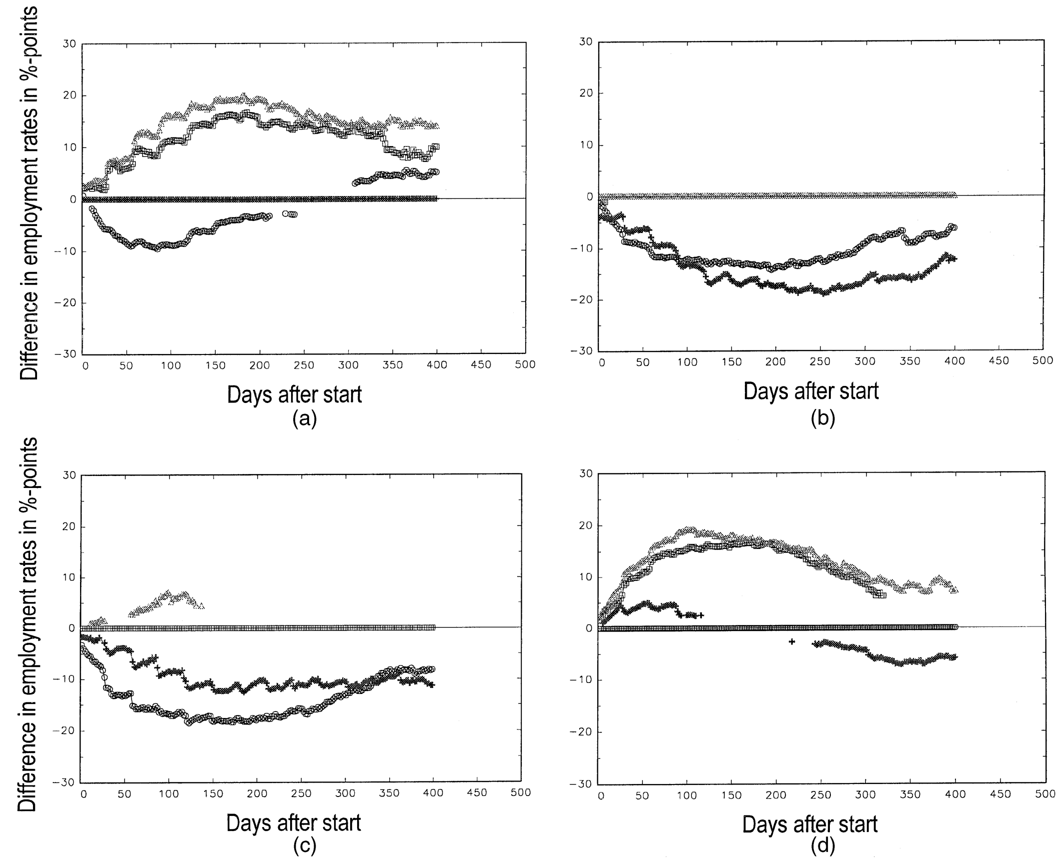

**Fig. 1.** Dynamics of average effects for participants after the start of the programme (only estimated effects that are significant at the 5% level are reported; ○, NONP; □, COC; △, EP; +, TEMP): (a) temporary wage subsidy; (b) employment programme; (c) computer course; (d) no programme

to cover several other issues that could be potentially responsible for the results obtained in the study by Gerfin and Lechner (2000). In addition to these the sensitivity of the results with respect to the amount of information included in the estimation will be addressed.

The various topics are structured in the following way. In Section 5.1 some fundamental specification problems that are directly related to identification are discussed. Section 5.2 is devoted to issues that could be considered as being technical relating to the implementation of the estimator and to obtaining valid inferences.

## 5.1. Fundamental issues

### 5.1.1. *Unknown start date of counterfactual programme*

Most ALMPs have the feature that individuals enter the various programmes at different times. Here, entries into the first programme are stretched over a period of 13 months (from January 2nd, 1998, to January 31st, 1999); however, about half of the entries are observed in the first quarter of 1998. The information about the start of the programme plays a role in two respects. First, it is used directly in the first step of the estimation (the multinomial probit model) and to compute several variables, like the duration of unemployment before the programme, that are assumed to be important in affecting participation in the programme and outcomes. Thus they are important to achieve identification. Second, the effect of the programmes is measured after their start.

There is a decision to be made about how to use or generate start dates. This decision obviously concerns non-participants, but in principle it is also relevant for participants of other programmes. The question is always 'when would the comparison person have started the programme?'. In the absence of any better hypothesis for participants, it is natural to assume that the start date is actually independent of the specific programme that the person is allocated to. In this case the observed start date could be used as a counterfactual start date for all the other programmes. If the start date is also independent of the characteristics of the individual, a natural choice for the participants is a random draw from the distribution of the observed start dates of all participants. For the binary treatment framework, other alternatives are discussed in Lechner (1999) that are applicable here as well. However, mainly because of their additional complexity they are less attractive in a multiprogramme evaluation that is more computer intensive than in a binary evaluation. Of course this procedure needs another adjustment for the case when the simulated start date is in contradiction to the administrative arrangements (here, an individual needs to be unemployed to enter a programme). In the base-line scenario this approach is used and the data for 'contradictory' non-participants, i.e. those with on average shorter unemployment spells (37% of all non-participants), have been deleted from the sample.

Although in specific applications the assumption of random start dates could be plausible, it is probably more plausible to assume that start dates could be predicted by the variables influencing outcomes and selection (as long as they do not depend on the start date). Again, in this case, using the observed start dates for the participants seems to be the best choice. For the non-participants start dates should be drawn from the conditional distribution of start dates given the covariates. As a sensitivity check, the logarithm of the start dates (the earliest day is 2; the latest is 391) are regressed on covariates, with start-date-dependent covariates substituted by proxies (the actual duration of unemployment is approximated by unemployed duration at the end of 1997, for example). To simulate the start date a log-normal distribution is assumed for the start day on the basis of a linear specification of its conditional mean (taken from the regression). It turns out that start dates can to some extent be predicted by using these covariates, although an $R^2$-value of 5% shows the limited amount of useful information that is contained in the covariates with respect to the timing of the programmes. The number of observations deleted reduces to 28%. In another check, this approach is used on a subsample of participants who enter the programme only in the first quarter of 1998, thus making the start date distribution more homogeneous. In this case the reduction of the sample of participants resulted in a loss of 50% of the participants. Only 12% of the data for the non-participants have been deleted.

To avoid flooding the reader with numbers Table 7 shows only the effects of NONP for non-participants, because they should be most sensitive to these changes in the specification. It appears that despite the considerable reduction in sample size in the final specification the sensitivity to these variations in the specification is small. This is confirmed by checking the dynamic patterns (Fig. 2). No substantial differences can be discovered, other than an increased variance due to the smaller sampler.

**Table 7.** Average effects of NONP for non-participants ($\theta_0^{\mathrm{NONP},l}$) 1 year after start: start dates for non-participants†

| | *Average effects (percentage points) for the following groups:* | | |
|---|---|---|---|
| | *COC* | *EP* | *TEMP* |
| Base-line | 2.1 (3.2) | **7.2** (2.3) | **−6.4** (1.6) |
| Predicted with covariates | 2.5 (2.9) | **8.5** (2.5) | **−4.2** (1.5) |
| Predicted with covariates and reduced sample | 2.9 (3.2) | **8.8** (3.0) | **−5.2** (1.7) |

†Standard errors are given in parentheses. Results are based on matched samples. Numbers in bold indicate significance at the 1% level (two-sided test).

### *5.1.2. Available information*

The data used for the empirical study are exceptional in that they contain rich information about the current spell of unemployment and previous employment histories. It is argued that such informative data are necessary to make the CIA a valid identifying assumption. In this subsection we check how sensitive the results are with respect to that information. In addition to the base-line specification, the following specifications are considered (note that each specification is less informative than the previous one):

(a) no long-term history—no information from the pension system about the last 10 years;
(b) no information on the duration of the current spell of unemployment;
(c) no subjective information—no subjective information on chances of employment as given by the case-worker;
(d) no information on the current spell of unemployment;
(e) no information about previous employment, skills and occupation;
(f) no regional information;

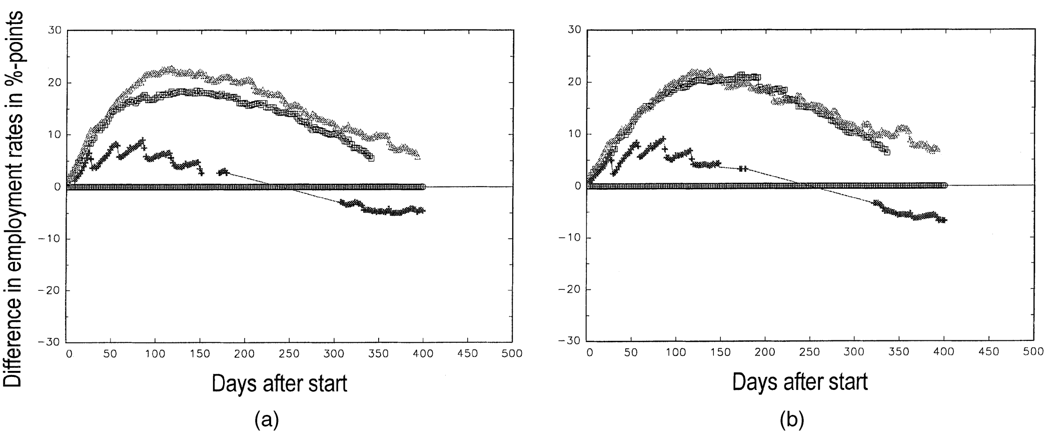

**Fig. 2.** Dynamics of average effects of NONP for non-participants $\theta_0^{\mathrm{NP},l}$ (only estimated effects that are significant at the 5% level are reported; ○, NONP; □, COC; △, EP; +, TEMP; for the base-line see Fig. 1(d)): (a) predicted with covariates; (b) predicted with covariates in the reduced sample

(g) only age, gender and marital status (no information on language and citizenship);
(h) no information (unadjusted differences).

Table 8 shows the effects for different specifications for one particular set of pairwise effects, namely the effects of COC for participants in such courses. *A priori* we would expect to see the most substantial changes here, because the participants appear to be clearly a positive selection in terms of unemployment risk, in particular compared with EP participants.

The results are indeed sensitive to shrinking the information set. Let us first consider the effects of COC compared with EP. Initially there is a small negative effect of COC that is insignificant, however. By removing information about the individual work-related characteristics the effect increases monotonically up to a level of 15%. It is only the removal of the regional information that does not change the estimates (conditional on the information that is available in the previous step). So, obviously, COC and EP participants have different chances in the labour market and any estimate of the effects needs to take account of these differences to avoid substantial biases in the estimated effects.

For the comparisons of COC with NONP and with TEMP—both programmes have less pronounced differences in the attributes of its participants compared with COC—the changes can be substantial but they are not necessarily monotonous, suggesting that in this case it is not necessarily 'better' to control for more variables than for 'fewer'.

The results from Table 8 are confirmed by considering the dynamics in Fig. 3. Although the patterns in all comparisons change, it is again the comparison between COC and EP that exhibits the largest effect.

Finally, a remark is in order with respect to the information that is contained in the subjective valuation of the labour offices. The changes in the estimate suggest that this information may indeed be valuable in uncovering characteristics that would otherwise

**Table 8.** Average effects of COC for participants ($\theta_0^{COC,l}$) 1 year after start: reduction of information†

| | *Average effects (percentage points) for the following groups:* | | |
|---|---|---|---|
| | *NONP* | *EP* | *TEMP* |
| Base-line | **−8.3** (2.5) | −2.1 (3.5) | **−9.1** (2.7) |
| and no long-term employment history | **−7.8** (2.5) | 1.0 (3.4) | **−8.8** (2.7) |
| and no duration of current spell of unemployment | **−8.9** (2.5) | 4.8 (3.3) | **−7.0** (2.7) |
| and no subjective information | *−5.0* (2.5) | *7.1* (3.3) | **−9.3** (2.7) |
| and no information on current spell of unemployment | −4.1 (2.5) | *7.9* (3.2) | **−8.8** (2.7) |
| and no information on previous employment, occupation and skill | 1.4 (2.5) | **14.1** (3.1) | **−10.5** (2.6) |
| and no regional information | −4.6 (2.5) | **14.1** (3.0) | −5.1 (2.7) |
| Only age, gender and marital status (no nationality) | 3.9 (2.4) | **14.7** (2.8) | **−9.7** (2.2) |
| No covariates (unadjusted differences) | **5.2** (1.7) | **15.0** (2.1) | *−4.2* (1.9) |

†Standard errors are given in parentheses. Results are based on matched samples. Numbers in bold indicate significance at the 1% level (two-sided test); numbers in italics indicate significance at the 5% level.

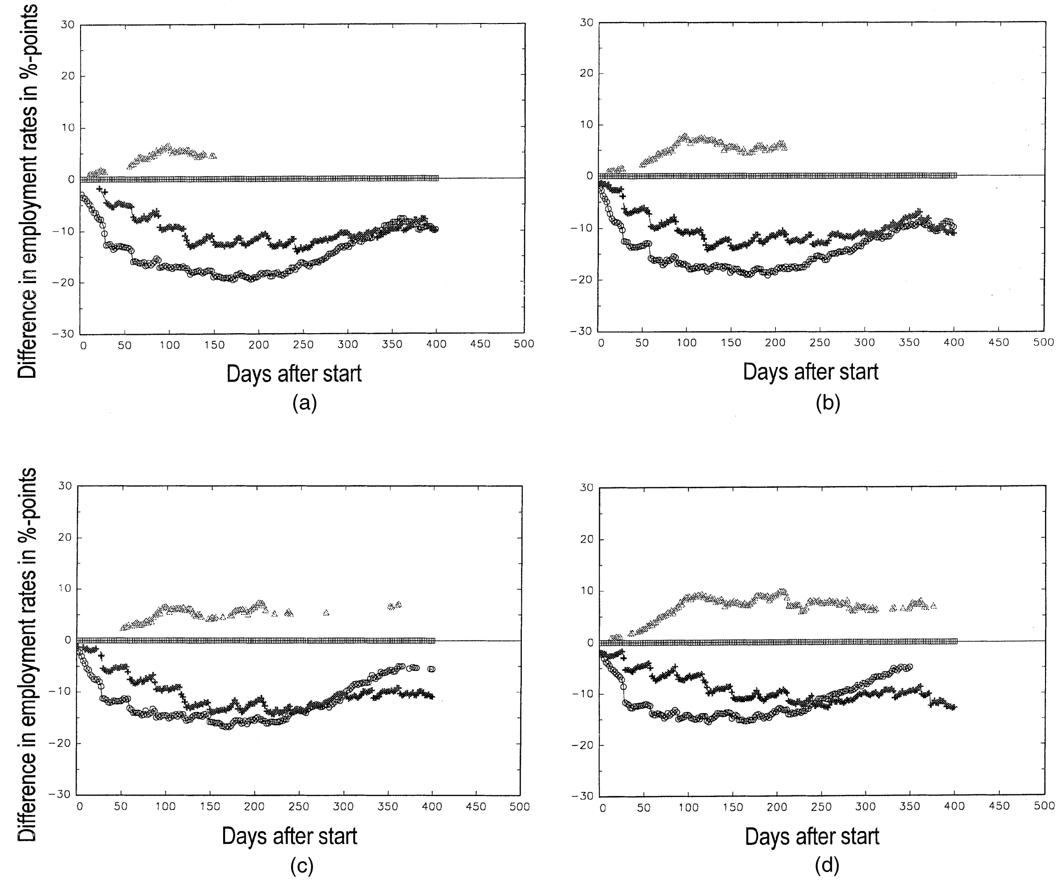

**Fig. 3.** Dynamics of average effects of COC for participants $(\theta_0^{COC,l})$—reduction of information (only estimated effects that are significant at the 5% level are reported; ○, NONP; □, COC; △, EP; +, TEMP; for the base-line see Fig. 1(c)): base-line and (a) no long-term employment history and (b) no duration of current spell of unemployment and (c) no subjective information and (d) no information on current spell of unemployment and (e) no information on previous employment and (f) no regional information; (g) only age, gender and marital status; (h) no matching

be left undetected (of course this observation is conditional on the information set used here).

## 5.2. Technical issues

### 5.2.1. Issues related to the first step of the estimation

The specification of the conditional probabilities could also have an influence on the results. The first decision to make is whether the conditional participation probabilities should be estimated for each combination of states separately as binary choices, or whether the process should be modelled simultaneously with a discrete choice model including all relevant states. The former has the advantage of being a more flexible specification, whereas the latter is much easier to monitor and to interpret. Lechner (2001a) devoted considerable attention to this problem and found that for a very similar application nothing was gained by going the

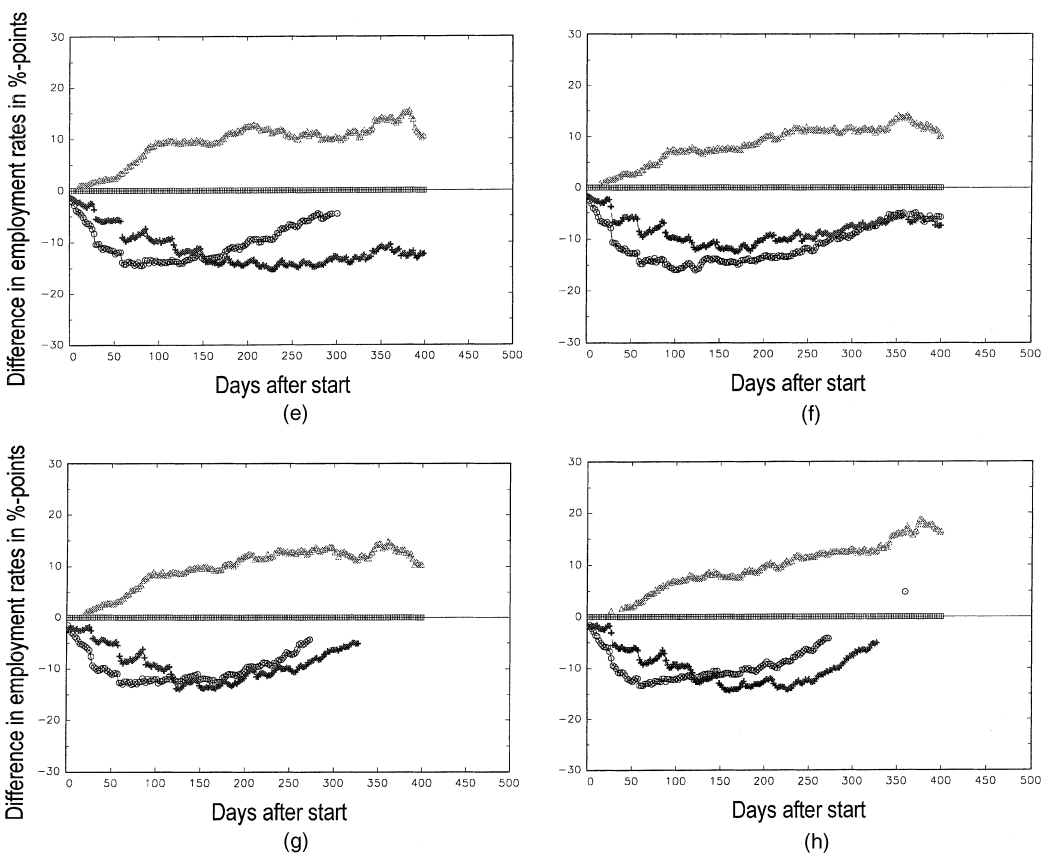

**Fig. 3** (*continued*)

more flexible route of modelling the binary choices separately. When using a multinomial discrete choice model a flexible version appears to be desirable. However, the computational costs may be substantial. The multinomial probit model estimated by simulated maximum likelihood is an attractive compromise, because it is sufficiently fast to compute but does not impose the so-called independence of irrelevant alternatives assumption, which the multinomial logit model does.

To check the sensitivity of the results with respect to the specification of the covariance matrix of the error terms appearing in the multinomial probit model choice equations, the covariance between the error terms of COC and all other alternatives are set to zero. Furthermore, the sensitivity of the results with respect to the number of simulations used in the GHK simulator is checked by computing the results for just two draws as well as 800 draws, whereas the base-line specification is based on 200 draws per choice equation and observation.

Again, since the results for COC could be expected to be most sensitive to those changes, they are presented in Table 9 and Fig. 4. From the results concerning the number of draws these issues do not appear to matter at all, because all changes are of the order of less than half a standard deviation of the estimator. The sensitivity with respect to the covariance structure is larger, however (more than 1 standard deviation in the comparison with NONP).

**Table 9.** Average effects of COC for participants ($\theta_0^{\mathrm{COC},l}$) 1 year after start: first step†

| | *Average effects (percentage points) for the following groups:* | | |
|---|---|---|---|
| | *NONP* | *EP* | *TEMP* |
| Base-line (200 draws, all 3 correlations between programmes) | −8.3 (2.5) | −2.1 (3.5) | −9.1 (2.7) |
| 2 draws | −8.5 (2.5) | 0.9 (3.5) | −8.3 (2.7) |
| 800 draws | −9.6 (2.5) | −0.8 (3.6) | −9.2 (2.7) |
| 3-way correlation between NONP and (TEMP, EP, COMP) | −9.1 (2.5) | −1.9 (3.5) | −9.1 (2.7) |
| Only correlation between TEMP and EP | −5.3 (2.6) | −0.9 (3.5) | −10.3 (2.7) |
| Only correlation between COMP and EP | −6.9 (2.5) | −1.9 (3.5) | −11.5 (2.7) |
| Only correlation between COMP and TEMP | −3.3 (2.6) | −2.3 (3.5) | −9.7 (2.7) |

†Standard errors are given in parentheses. Results are based on matched samples.

On the one hand this finding suggests that using a discrete choice model that relies on a more restrictive specification, like the multinomial logit model, could lead to biases. On the other hand, there could be an argument for avoiding multinomial models altogether and using (many) binary models instead.

### *5.2.2. The common support requirement*

The CIA implies that the decision to participate can be considered as random conditional on the covariates. To be non-trivial 'randomness' requires that for a given vector of covariates there is a positive probability of participating in every programme. The first step to ensure that this requirement is satisfied in an application is to consider only individuals who—according to the institutional settings—could in principle participate in the programmes under consideration. In the current study this refers to the requirement that individuals had to be unemployed on December 31st, 1997 (in addition to some other requirements; see Gerfin and Lechner (2000)). As a property of a multinomial probit model the estimated conditional probabilities for all individuals are strictly bounded away from zero. However, we may find (extreme) values of the covariates that generate conditional probabilities for participants in one programme that cannot be found for participants in other programmes. Hence, there is no way to estimate the effect for this (extreme) group with the sample at hand. At this point there are two ways to proceed. The obvious way is to ignore this problem by referring to asymptotics: although the probabilities of being observed in a particular state with such covariates may be very small, eventually (which means with some other random sample) there will be such an observation and matching will be satisfactory. Of course, with the data at hand there will be a (finite sample) bias if the potential outcomes vary with the probabilities, because these (extreme) cases lead to bad matches. The second option is to ensure that the distributions of the balancing scores overlap by removing extreme cases. The drawback here is that the definition of the treatment effects are changed in the sense that they are now mean effects for a narrower population defined by the overlap in the support.

Table 4 already showed the loss of observations when restricting the sample by considering the smallest maximum and the largest minimum in the subsamples as joint

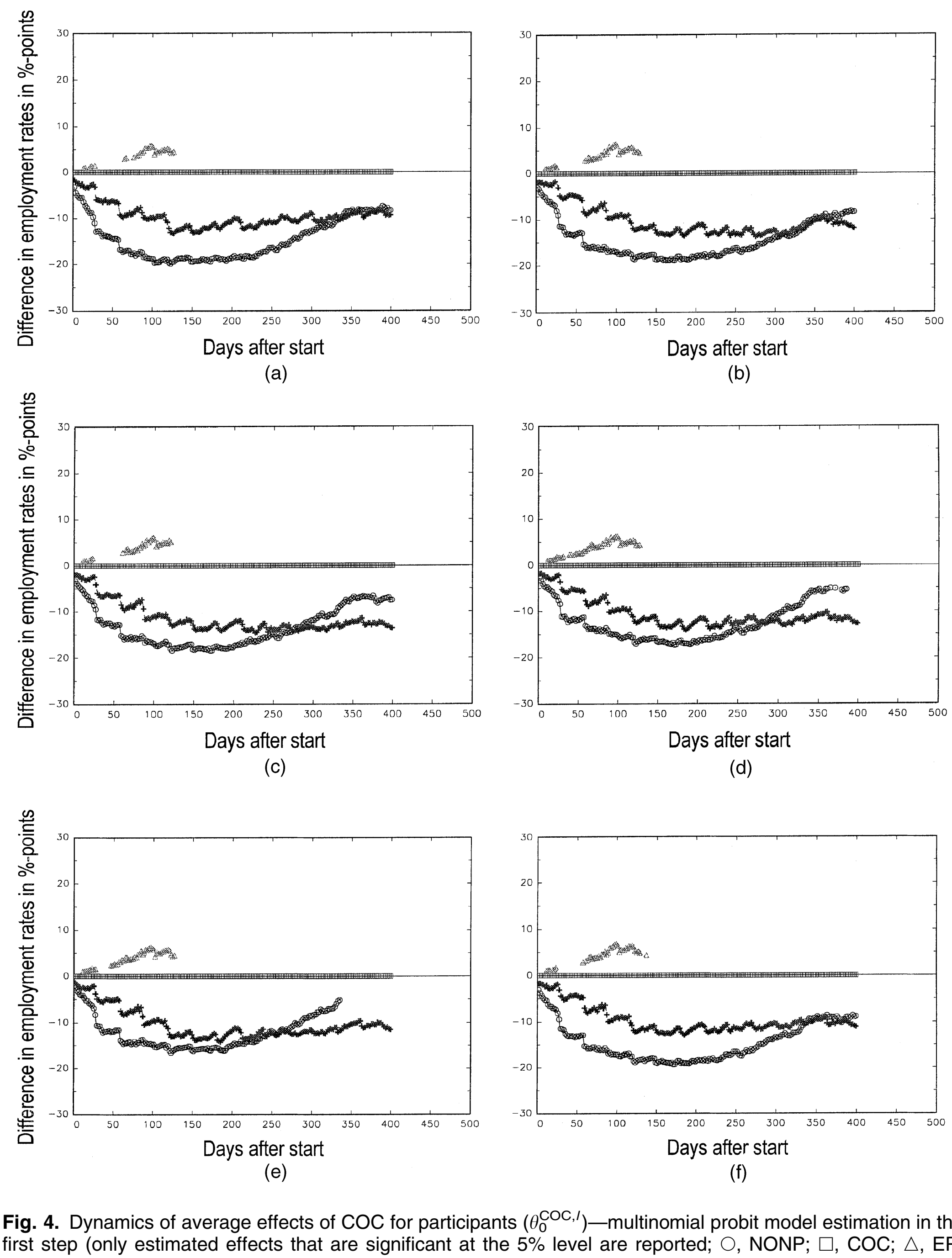

**Fig. 4.** Dynamics of average effects of COC for participants ($\theta_0^{COC,l}$)—multinomial probit model estimation in the first step (only estimated effects that are significant at the 5% level are reported; ○, NONP; □, COC; △, EP; +, TEMP; for the base-line see Fig. 1(c)): (a) two draws in the GHK simulator; (b) 800 draws in the GHK simulator; (c) correlation between COC and EP; (d) correlation between TEMP and EP; (e) correlation between COC and TEMP; (f) mutual correlations between NONP and COC, TEMP and EP

**Table 10.** Average effects of TEMP for participants ($\theta_0^{\text{TEMP},l}$) 1 year after start: first step†

| | *Average effects (percentage points) for the following groups:* | | |
|---|---|---|---|
| | *NONP* | *COC* | *EP* |
| Base-line | 4.2 (1.7) | 8.0 (3.3) | 13.8 (2.7) |
| No common support | 2.3 (1.7) | 7.5 (3.3) | 13.5 (2.7) |
| Stricter common support requirement | 3.9 (1.8) | 8.1 (3.3) | 14.9 (2.7) |

†Standard errors are given in parentheses. Results are based on matched samples.

bounds for the common support. The overall loss of observations is rather small. One could argue that the density in the tail of the implied distributions is still very thin, because there could be a substantial distance for example from the smallest maximum to the second smallest element of that probability in this specific subsample. Therefore, to check the sensitivity a stricter requirement is imposed, where the maximum and the minimum are substituted by the 10th largest and 10th smallest observations. The suspicion that the density may be thin seems to be justified, because the number of observations that are lost due to that more restrictive requirement increases from about 1–3% (see Table 4) to 16% for NONP, 14% for COC, 15% for EP and 19% for TEMP. Because TEMP seems to be most affected by these changes, Table 10 as well as Fig. 5 show the effects for this programme.

When the common support condition is not enforced, the major change is that the positive effect with respect to NONP is reduced and is no longer significant at the 5% level, which indeed changes one important policy conclusion. Another change concerns the increased effect in comparison with EP. However, this increase by 1 percentage point is less than half a

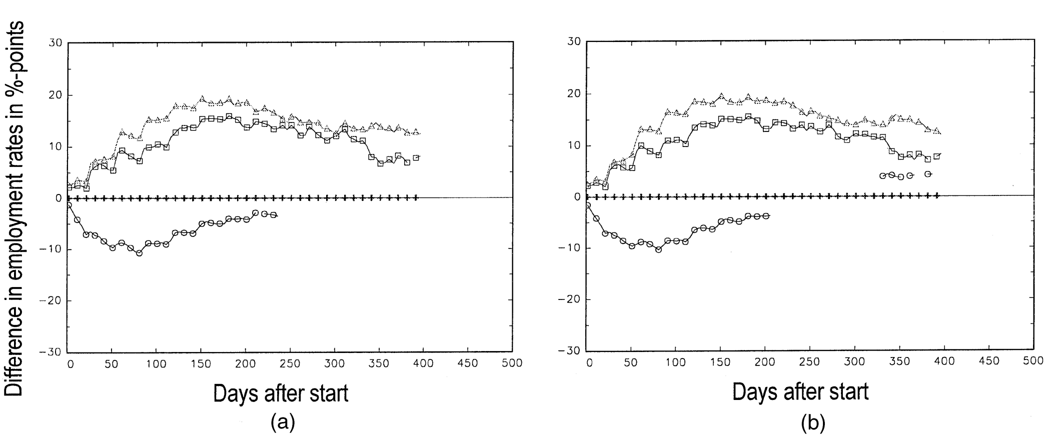

**Fig. 5.** Dynamics of average effects of TEMP for participants ($\theta_0^{\text{TEMP},l}$)—common support (only estimated effects that are significant at the 5% level are reported; ○, NONP; □, COC; △, EP; +, TEMP; for the base-line see Fig. 1(a)): (a) no common support required; (b) stricter common support required

standard deviation of the estimator and hence it is not substantial. To summarize, these results tend to suggest the importance of removing extreme observations. Since matching is with replacement and the samples are large, the additional trimming of thin tails seems not to be necessary, at least in this application.

Lechner (2001c) suggests another way in addition to the conventional removal of observations. The idea entertained there is that, although the original effect of interest is not identified without common support, the information that is available may nevertheless be used to obtain sharp bounds in cases when the expectation of the outcome variable is finite with known lower and upper limits.

### *5.2.3. Asymptotic distribution*

This study has so far conducted inference based on the presumption that the estimators have an asymptotic normal distribution derived from the difference of two weighted means of independent observations. This approximation, however, ignores the fact that the comparison group is formed by matching using an estimated balancing score based on a simulation of start dates for non-participants. Furthermore, estimated probabilities are used for the data-driven reduction of the sample to ensure the common support criterion. So far no asymptotic theory taking account of these features of the estimator has been developed. One way to check the accuracy of this approximation for the current study is to compare the approximation with an inference based on bootstrapping. Since each estimation is fairly expensive in terms of computation time, the bootstrap is based on only 400 bootstrap samples. For each estimation a new sample of the same size is drawn with replacement and all the steps of the estimation, including simulation of start dates and the enforcement of common support, are performed on the simulated sample.

Table 11 compares several estimates obtained from the bootstrap samples with those obtained from the approximation. Quite arbitrarily the results are given only for TEMP. However, the other results are similar. Table 11 displays the results for the mean, the

**Table 11.** Average effects of TEMP for participants ($\hat{\theta}_0^{\text{TEMP},l}$) 1 year after start: bootstrap†

| *Group* | $\hat{\theta}_N^{\text{TEMP},l}$ | *Standard deviation* | *Average effects (percentage points) for the following quantiles:* | | | | | | | *Normality p-value × 100* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *2.5%* | *5%* | *25%* | *Median* | *75%* | *95%* | *97.5%* | |
| *Approximation* | | | | | | | | | | |
| NONP | 4.2 | 1.7 | 0.9 | 1.4 | 3.0 | 4.2 | 5.4 | 7.0 | 7.5 | |
| COC | 8.0 | 3.3 | 1.5 | 2.6 | 5.8 | 8.0 | 10.2 | 13.4 | 14.5 | |
| EP | 13.8 | 2.7 | 8.5 | 9.4 | 12.0 | 13.8 | 15.6 | 18.2 | 19.1 | |
| *Bootstrap* | | | | | | | | | | |
| NONP | 4.3‡ | 2.0§ | 0.0 | 1.2 | 3.1 | 4.2 | 5.8 | 7.6 | 8.4 | 4 |
| COC | 8.0‡ | 3.5§ | 1.1 | 1.7 | 5.6 | 8.1 | 10.4 | 13.5 | 14.2 | 32 |
| EP | 13.8‡ | 2.9§ | 8.3 | 8.9 | 11.9 | 14.0 | 15.8 | 19.0 | 19.2 | 17 |

†Results are based on matched samples. 400 bootstrap samples. The bootstrap quantiles are based on the empirical order statistic ($\hat{\theta}_{N,h}^{\text{TEMP},l}$). Normality is tested by the skewness–kurtosis statistic that is asymptotically distributed as $\chi^2(2)$ and attributed to Fisher (see for example Spanos (1999), page 745).
‡Mean $\hat{\theta}_{N,h}^{\text{TEMP},l}$.
§Standard deviation $\hat{\theta}_{N,h}^{\text{TEMP},l}$.

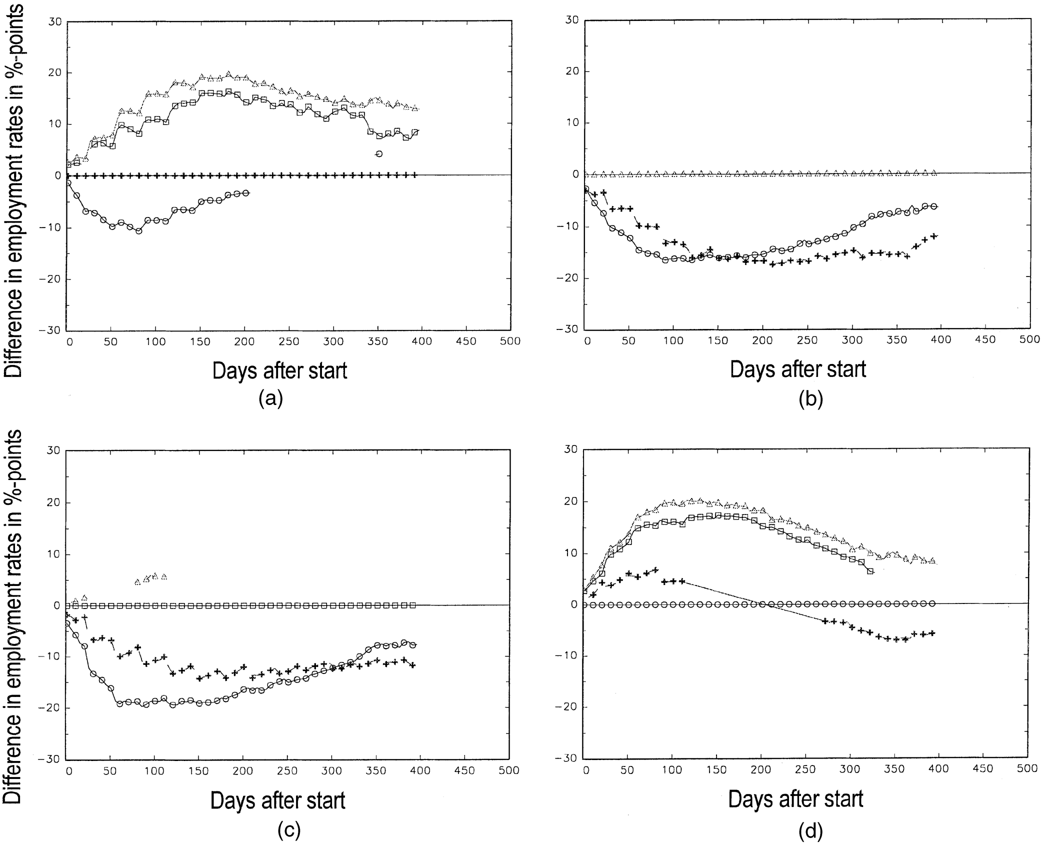

**Fig. 6.** Dynamics of average effects for participants after the start of the programme—bootstrap results based on 400 samples (only estimated effects that are significant at the 5% level are reported and only every fifth day is displayed; effects are only displayed if the bootstrap bounds of the 95% interval have the same sign; ○, NONP; □, COC; △, EP; +, TEMP): (a) temporary wage subsidy; (b) employment programme; (c) computer course; (d) no programme

standard deviation and some quantiles that are commonly used in inference. Since the more extreme quantiles could be subject to considerable simulation error owing to the small number of bootstrap replications, the 25% and the 75% quantile are given as well. In addition Fisher's test for normality based on the skewness and kurtosis of the distribution of the effect across the bootstrap samples is shown. It turns out that the results based on the approximation and those based on the bootstrap are fairly similar. There is probably a slight underestimation of the variability of the estimates by the approximation.

Fig. 6 presents the corresponding dynamics for all the treatments. An effect is only displayed if the upper and lower bound of the 95% empirical bootstrap interval have the same sign. It is very difficult to spot any difference between Fig. 1 and the bootstrap results. Thus the base-line results are again confirmed. Given the computer intensiveness of the bootstrap for large samples, the approximation has a considerable attraction. However, the usual pace of development in computer technology may change that observation in the future.

## 6. Conclusion

The study by Gerfin and Lechner (2000) analysed the Swiss ALMPs by using a newly proposed matching estimator for multiple programmes. The study is based on rich data, so conditioning on the information that is available in that data, selection for the various programmes and the outcome variables are probably mutually independent. Furthermore, the sample sizes are comparatively large.

In such a situation the matching estimator in its multiple programme version is an attractive choice. It has the advantage that it is basically nonparametric or at least semiparametric so very few additional assumptions are necessary at the estimation stage of the analysis. Furthermore, it allows the effect to vary across individuals and programmes in an unrestricted way. Finally, the principles underlying this estimator are fairly easy to communicate to non-statistical users of evaluation studies.

There is only a very limited practical experience with these kinds of matching estimator for multiple programmes. In this paper the sensitivity of this estimator with respect to some features that are of importance in empirical studies has been checked. It turns out that the estimator is fairly robust to several issues that concern its implementation. The only exception to some extent is the specification of the probability model that is used to predict the various participation probabilities that form the basis for matching. The comparison with a bootstrap distribution provides some justification for the common use of a simplified approximation of the distribution of the matching estimator that ignores several issues relating to its sequential nature.

The paper also demonstrates that the matching approach *per se* is no panacea for solving all the problems of evaluation studies, but that its success depends critically on the information that is available in the data, i.e. whether using the CIA for identification is plausible. Given the obvious insight that the performance of this estimator depends on the information that is available, any discussion about whether this or any other estimator is the 'best' estimator for evaluation studies in general is obviously misguided.

Although matching cannot solve all the potential problems of an evaluation study, if identification can be achieved by rich data and sufficient institutional knowledge about the selection process, then it is the opinion of the author that some version of matching is clearly the estimator of choice. However, if the CIA is not plausible, then there is no *a priori* reason why matching should be any better than any other evaluation estimator. In this case the researcher must decide whether to collect more data or to find another plausible identifying assumption.

## Acknowledgements

I am also affiliated to the Centre for Economic Policy Research, London, Zentrum für Europäische Wirtschaftsforschung, Mannheim, and IZA, Bonn. Financial support from the Swiss National Science Foundation (grant NFP 12-53735.18) is gratefully acknowledged. The data are a subsample from a database prepared for the evaluation of the Swiss ALMP together with Michael Gerfin. I am grateful to the State Secretariat of Economic Affairs of the Swiss Government and the Bundesamt für Sozialversicherung for providing the data and to Michael Gerfin for his help in preparing them. I also thank Markus Frölich for comments and for providing me with an improved version of the Gauss code for the multinomial probit model. Helpful comments from the Joint Editor and two referees of this journal are gratefully acknowledged. Any remaining errors are my own.

## References

Angrist, J. D. (1998) Estimating labor market impact of voluntary military service using social security data. *Econometrica*, **66**, 249–288.

Angrist, J. D. and Krueger, A. B. (1999) Empirical strategies in labor economics. In *Handbook of Labor Economics* (eds O. Ashenfelter and D. Card), vol. IIIA, ch. 23, pp. 1277–1366. Amsterdam: North-Holland.

Börsch-Supan, A. and Hajivassiliou, V. A. (1993) Smooth unbiased multivariate probabilities simulators for maximum likelihood estimation of limited dependent variable models. *J. Econometr.*, **58**, 347–368.

Brodaty, T., Crepon, B. and Fougère, D. (2001) Using matching estimators to evaluate alternative youth employment programmes: evidence from France, 1986-1988. In *Econometric Evaluation of Labour Market Policies* (eds M. Lechner and F. Pfeiffer), pp. 85–123. Heidelberg: Physica.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in non-experimental studies: reevaluating the evaluation of training programmes. *J. Am. Statist. Ass.*, **94**, 1053–1062.

Dorsett, R. (2001) The New Deal for Young People: relative effectiveness of the options. *Mimeo*. Policy Studies Institute, London.

Fay, R. G. (1996) Enhancing the effectiveness of active labour market policies: evidence from programme evaluations in OECD countries. *Labour Market and Social Policy Occasional Paper 18*. Organisation for Economic Co-operation and Development, Paris.

Frölich, M., Heshmati, A. and Lechner, M. (2000) A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. *Discussion Paper 2000-04*. University of St Gallen, St Gallen.

Gerfin, M. and Lechner, M. (2000) Microeconometric evaluation of the active labour market policy in Switzerland. *Discussion Paper 2000-10*. University of St Gallen, St Gallen.

Geweke, J., Keane, M. and Runkle, D. (1994) Alternative computational approaches to inference in the multinomial probit model. *Rev. Econ. Statist.*, **76**, 609–632.

Heckman, J. J., Ichimura, H., Smith, J. A. and Todd, P. (1998) Characterisation selection bias using experimental data. *Econometrica*, **66**, 1017–1098.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Rev. Econ. Stud.*, **64**, 605–654.

——— (1998) Matching as an econometric evaluation estimator. *Rev. Econ. Stud.*, **65**, 261–294.

Heckman, J. J., LaLonde, R. J. and Smith, J. A. (1999) The economics and econometrics of active labour market programs. In *Handbook of Labor Economics* (eds O. Ashenfelter and D. Card), vol. IIIA, ch. 31, pp. 1865–2097. Amsterdam: North-Holland.

Heckman, J. J. and Robb, R. (1986) Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inferences from Self-selected Samples* (ed. H. Wainer), pp. 63–107. New York: Springer.

Holland, P. W. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–970.

Imbens, G. W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**, 706–710.

Lalive, R., van Ours, J. C. and Zweimüller, J. (2000) The impact of active labor market policies and benefit entitlement rules on the duration of unemployment. *Mimeo.*

Larsson, L. (2000) Evaluation of Swedish youth labour market programmes. *Discussion Paper 2000:1*. Office for Labour Market Policy Evaluation, Uppsala.

Lechner, M. (1999) Earnings and employment effects of continuous off-the-job training in East Germany after unification. *J. Bus. Econ. Statist.*, **17**, 74–90.

——— (2000) An evaluation of public sector sponsored continuous vocational training programs in East Germany. *J. Hum. Resour.*, **35**, 347–375.

——— (2001a) Programme heterogeneity and propensity score matching: an application to the evaluation of active labour market policies. *Rev. Econ. Statist.*, to be published.

——— (2001b) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* (eds M. Lechner and F. Pfeiffer), pp. 43–58. Heidelberg: Physica.

——— (2001c) A note on the common support problem in applied evaluation studies. *Discussion Paper 2001-01*. Department of Economics, University of St Gallen, St Gallen.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–50.

Roy, A. D. (1951) Some thoughts on the distribution of earnings. *Oxf. Econ. Pap.*, **3**, 135–146.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Smith, J. and Todd, P. E. (2000) Does matching overcome Lalonde's critique of nonexperimental estimators. *J. Econometr.*, to be published.

Spanos, A. (1999) *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press.