Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies

Michael Lechner [*]

University of St. Gallen

Swiss Institute for International Economics and Applied Economic Research (SIAW)

First version: September 1999

Date this version has been printed: 06 January 2000

Comments welcome

Michael Lechner

Professor of Econometrics

Swiss Institute for International Economics and Applied Economic Research (SIAW)

University of St. Gallen

Dufourstr. 48, CH-9000 St. Gallen, Switzerland

Michael.Lechner@unisg.ch, www.siaw.unisg.ch/lechner

[*] Financial support from the Swiss National Science Foundation (NFP 12-53735.18) is gratefully acknowledged. The data are a subsample from a data base generated for the evaluation of the Swiss active labour market policy together with Michael Gerfin. I am grateful to the Department of Economics of the Swiss Government (*seco; Arbeitsmarktstatistik*) for providing the data and to Michael Gerfin for his help in preparing them. The paper has been presented at the workshop *Evaluation of Labour Market Policies*, Bundesanstalt für Arbeit (IAB), Nuremberg, 1999. I thank participants for helpful comments and suggestions. All remaining errors are mine.

Abstract

This paper investigates the question whether it really matters for microeconometric evaluation studies to take account of the fact that the programmes under consideration are heterogeneous. Assuming that selection into the different sub-programmes and the potential outcomes are independent given observable characteristics, estimators based on different propensity scores are compared and applied to the analysis of the active labour market policy in a Swiss region. Furthermore, the issues of heterogeneous effects and aggregation are addressed. The econometric considerations as well as the results of the application suggest that an approach that incorporates the possibility of having multiple programmes could be an important tool in applied work.

Keywords

Multiple programmes, programme evaluation, treatment effects, balancing score, matching.

JEL classification: C10, C50, J60, J68.

1 Introduction

With respect to programme heterogeneity there is a big discrepancy between technically sophisticated modern microeconometric evaluation methods and real programmes to be evaluated. Standard microeconometric evaluation methods are mainly concerned with the effects of being or not being in a particular programme, whereas for example in active labour market policies (ALMP) typically there is a range of different versions of heterogeneous sub-programmes, such as training, public employment programmes, or job counselling.[1] These sub-programmes often differ with respect to their target population, their contents and duration, their selection rules as well as with respect to their effects.

For the case in which the participation in such a programme is independent of the subsequent outcomes conditionally on observable exogenous factors (conditional independence assumption, CIA) the standard model of only two states, i.e. participation versus nonparticipation, is extended by Imbens (1999) and Lechner (1999b) to the case of multiple states ('treatments').[2] Both papers show that the important dimension reducing device in the binary treatment model, called the balancing score property of the propensity score, is still valid in principle, but needs to be suitable revised.

This paper extends Lechner (1999b) in several aspects relating to the definition of treatment effects and the issue of aggregating different states. Furthermore, several estimation methods, all based on 'matching on the propensity score', are proposed. These methods have been applied to the evaluation of active labour market policies in the Swiss canton of Zurich. The aim of this study is to give an example of how an evaluation could be performed in this setting, and not to derive policy relevant conclusions. The comparison of different estimators in practise provides information about their practical performance. In addition, the application showed that the multiple treatment approach can lead to valuable insights that might be lost otherwise.

The next section defines the concept of causality, introduces the necessary notation and discusses identification of different effects for the case of multiple treatments based on the conditional

[1] For recent surveys of this literature see for example Angrist and Krueger (1999) and Heckman, LaLonde, and Smith (1999). The reader should note that in several previous studies the author of this paper ignored the existence of other programmes as well, thus being subject to the same criticism that will be brought forward in this paper.

[2] Note that the term multiple treatments also includes the issue of dose-response, since for example an employment programme with two different durations (the doses) could always be redefined as being two different programmes.

independence assumption. It also defines a causal effect corresponding to the aggregation of the different types of nonparticipation. Section 3 proposes matching estimators for this setting. Section 4 presents empirical results for the Swiss region of Zurich and Section 5 concludes. An Appendix contains some technical details.

2 The causal evaluation model with multiple treatments

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, like participation in a training programme or non-participation in such a programme. The potential participant in a programme will get an hypothetical outcome in both states. This model is known as the Roy (1951) - Rubin (1974) model of potential outcomes and causal effects.[3] Since its statistical content is most clearly spelled out in Rubin (1974), for simplicity this model is called the Rubin-model in the following.

Consider the outcomes of (*M+1*) different mutually exclusive states denoted by $\{Y^0, Y^1, ..., Y^M\}$. The different *states* will to be called *treatments* in the following to stick to the terminology of that literature. It is assumed that each participant receives exactly one of the treatments (typically, category '0' denotes treatment type *no treatment*). Therefore, for any participant, only one component of $\{Y^0, Y^1, ..., Y^M\}$ can be observed in the data. The remaining *M* outcomes are counterfactuals in the language of the Rubin model. Participation in a particular treatment *m* is indicated by the variable $S \in \{0,1,...M\}$. The number of observed participants in treatment *m* is denoted by N^m ($N = \sum_{m=0}^{M} N^m$).

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

2.1.2 Pair-wise effects

The definitions of average treatment effects used for the case of just two treatments need to be extended.[4] In the following equations, the focus is on a pair-wise comparison of the effects of treatments *m* and *l*:

$$\gamma_0^{m,l} = E(Y^m - Y^l) = EY^m - EY^l; \quad (1)$$

$$\alpha_0^{m,l} = E(Y^m - Y^l | S = m,l) = E(Y^m | S = m,l) - E(Y^l | S = m,l); \quad (2)$$

$$\theta_0^{m,l} = E(Y^m - Y^l | S = m) = E(Y^m | S = m) - E(Y^l | S = m). \quad (3)$$

$\gamma_0^{m,l}$ denotes the expected (average) effect of treatment *m* relative to treatment *l* for a participant drawn randomly from the population.[5] Similarly, $\alpha_0^{m,l}$ denotes the same effect for a participant randomly selected from the group of participants participating in either *m* or *l*. Note that both average treatment effects are symmetric in the sense that $\gamma_0^{m,l} = -\gamma_0^{l,m}$ and $\alpha_0^{m,l} = -\alpha_0^{l,m}$.[6] $\theta_0^{m,l}$ is the expected effect for an individual randomly drawn from the population of participants in treatment *m* only. Note that if the participants in treatments *m* and *l* differ in a way that is related to the distribution of *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.

It is worth noting that $\alpha_0^{m,l} = E(Y^m - Y^l | S = m,l)$ is a weighted combination of $\theta_0^{m,l}$ and $\theta_0^{l,m}$. The weights are given by the participation probabilities in the respective states *m* and *l* (see Lechner, 1999b):

$$\alpha_0^{m,l} = \theta_0^{m,l} P(S = m | S = m,l) - \theta_0^{l,m}[1 - P(S = m | S = m,l)]. \quad (4)$$

---

[4] Assume for the rest of the paper that the typical assumptions of the Rubin model are fulfilled (see Holland, 1986, or Rubin, 1974, for example).

[5] If a variable *Z* cannot be changed by the effect of the treatment (like time constant personal characteristics of participants), then all what follows is also valid in strata of the data defined by different values of *Z*.

[6] For *m* = *l*, all effects are of course zero.

2.1.3 Composite effects

Since the pair-wise comparison may not be considered an optimal way to summarize the causal effects in the case of many treatments, the following modifications can be used to define a composite (or aggregate) effect by using appropriate weight functions to aggregate the treatments other than *m*:

$$\gamma_0^m(v^m) = \sum_{l=0}^{M} v^{m,l}\gamma_0^{m,l}, \qquad v^m = (v^{m,0},...,v^{m,M})'; \tag{5}$$

$$\alpha_0^m(v^m) = \sum_{l=0}^{M} v^{m,l}\alpha_0^{m,l}; \tag{6}$$

$$\theta_0^m(v^m) = \sum_{l=0}^{M} v^{m,l}\theta_0^{m,l}. \tag{7}$$

For a useful interpretation of these effects the weight functions should fulfil $v^{m,m} = 0$ and $\sum_{l=0}^{M} v^{m,l} = 1$. The above notation could be used for example to define treatment effects that are measured relative to some average of treatment outcomes other than those of treatment *m*. Obviously the pair-wise effects defined in equations (1) to (3) are special cases.

Although the composite effects given in equations (5) to (7) do not look like causal effects at first sight, $\gamma_0^m(v^m)$ and $\theta_0^m(v^m)$ have nevertheless a causal interpretation, since they correspond to the effects of treatment *m* compared to a state were the treated would be randomly assigned to one of the other treatments with probabilities given by the weights. Thus the composite potential outcome is defined $Y^{-m}(v^m) = \sum_{l=0}^{M} v^{m,l}\, Y^l$, where weights are given constants with $v^{m,m} = 0, \sum_{l=0}^{M} v^{m,l} = 1$. Then the composite effects can be rewritten as (proof in Appendix A):

$$\gamma_0^m(v^m) = E(Y^m) - E[Y^{-m}(v^m)]; \tag{8}$$

$$\theta_0^m(v^m) = E(Y^m \mid S = m) - E[Y^{-m}(v^m) \mid S = m]. \tag{9}$$

Unfortunately, such an interpretation is not possible for $\alpha_0^m(v^m)$, because the implicit conditioning set used to weight the various pair-wise effects depends on both potential outcomes appearing in each pair-wise effect $\alpha_0^{m,l}$ (see Appendix A).

2.2 The conditional independence assumption

The Rubin model clarifies that the average causal treatment effect - defined as the average difference of the two potential outcomes in some population, for example - is generally not identified. Therefore, the lack of identification has to be overcome by plausible, untestable assumptions that usually depend heavily on the problem analyzed and the data available. One such assumption is that treatment participation and treatment outcome is independent conditional on a set of (observable) attributes (conditional independence assumption, CIA). Subsequent papers by Rubin (1977) and Rosenbaum and Rubin (1983) show how this assumption could effectively be used for treatment evaluation since it is not necessary to condition on the attributes, but only the participation probability conditional on the attributes. In many cases this identifying assumption is exploited via a matching estimator, for recent examples see Angrist (1998), Heckman, Ichimura, and Todd (1998), and Lechner (1999a).

Imbens (1999) and Lechner (1999b) consider identification under the conditional independence assumption (CIA), that states that the potential treatment outcomes are independent of the assignment mechanism for any given value of a vector of attributes (X) in a particular attribute space χ. This assumption is formalised in expression (10):

$$Y^0, Y^1, \ldots, Y^M \coprod S | X = x, \forall x \in \chi. \qquad (10)$$

In an observational study it requires the researcher to observe all characteristics that jointly influence the outcomes as well as the selection into the treatments. In that sense, CIA may be called a 'data hungry' identification strategy. Note that CIA can be seen as overly restrictive, since all what is needed to identify mean effects is conditional mean independence. However, the former has the virtue of making the latter valid for all transformations of the outcome variables. Furthermore, in an application it is usually difficult to argue why conditional mean independence should hold and CIA might nevertheless be violated.

2.3 Identification and the balancing score

2.3.1 Pair-wise effects

This section discusses the identification of $\theta_0^{m,l}$ and $\gamma_0^{m,l}$ from an infinitely large random sample. In such a sample all participation probabilities are identified. There is no need to address the identification of $\alpha_0^{m,l}$ explicitly, because it is a weighted average of $\theta_0^{m,l}$ and $\theta_0^{l,m}$ (equation (4)). Thus, $\alpha_0^{m,l}$ is identified whenever $\theta_0^{m,l}$ and $\theta_0^{l,m}$ are identified.

Lechner (1999b) shows that CIA identifies all effects as long as each cell has a marginal probability conditional on *X* larger than *0* and smaller than *1*. Furthermore, that paper shows that some modified versions of the balancing score properties known from the binary treatment model (Rosenbaum and Rubin, 1983) hold in this more general setting as well. In the following the basic results of Lechner (1999b) are repeated.

Denote the choice probability of alternative *j* conditional on *X* as $P(S = j \mid X = x) = P^j(x)$, then for the pair-wise treatment effect in the population the following equation is obtained:

$$\gamma_0^{m,l} = E(Y^m \mid S = m)P(S = m) + \underset{P^m(X)}{E}[E(Y^m \mid P^m(X), S = m) \mid S \neq m]P(S \neq m)$$

$$- E(Y^l \mid S = l)P(S = l) + \underset{P^l(X)}{E}[E(Y^l \mid P^l(X), S = l) \mid S \neq l]P(S \neq l). \qquad (11)$$

If the respective probabilities $P^m(x)$ and $P^l(x)$ are known or if a good estimator is available, i.e. a consistent estimator that converges at the parametric rate, the dimension of the (nonparametric) estimation problem is reduced to one.

Lechner (1999b) gives a similar result for the pair-wise treatment effect for the participants in one of the treatments:

$$\theta_0^{m,l} = E(Y^m \mid S = m) + \underset{P^{l|ml}(X)}{E}[E(Y^l \mid P^{l|ml}(X), S = l) \mid S = m]. \qquad (12)$$

$$P^{l|ml}(x) = P^{l|ml}(S = l \mid S = l \text{ or } S = m, X = x) = \frac{P^l(x)}{P^l(x) + P^m(x)}.$$

Again, the dimension of the estimation problem is reduced to one. If it is possible to model $P^{l|ml}(x)$ directly, no information from subsamples other than the participants in *m* and *l* is needed for identification and hence for the estimation of $\theta_0^{m,l}$ and $\theta_0^{l,m}$. In many cases however, it will be more straightforward from a modelling point of view to model the complete discrete choice problem of choosing a particular treatment out of the complete list of treatments simultaneously. $P^{l|ml}(x)$ could then be computed from that model. When a discrete choice model is estimated, or generally when the conditional choice probabilities are more difficult to obtain than the marginal ones, it may be attractive to condition jointly on $P^l(X)$ and $P^m(X)$ instead of $P^{l|ml}(X)$. This also identifies $\theta_0^{m,l}$, because $P^l(X)$ together with $P^m(X)$ is finer than $P^{l|ml}(X)$, since

$$E[P^{l|ml}(X)\,|\,P^l(X),P^m(X)] = E[\frac{P^l(X)}{P^l(X)+P^m(X)}\,|\,P^l(X),P^m(X)] = P^{l|ml}(X).$$

The equality $E(Y^l \mid S=m) = E_X[E(Y^l \mid P^{l|ml}(X), S=l) \mid S=m]$ that is used in equation (12) suggests another way of identifying (and estimating) the population effect $\gamma_0^{m,l}$, because $\gamma_0^{m,l}$ can be written as follows:

$$\gamma_0^{m,l} = EY^m - EY^l = \sum_{j=0}^{M}[E(Y^m \mid S=j) - E(Y^l \mid S=j)]P(S=j)$$

$$= \sum_{j=0}^{M}\{E_X[E(Y^m \mid P^{m|mj}(X), S=m) \mid S=j] - E_X[E(Y^l \mid P^{l|lj}(X), S=l) \mid S=j]\}P(S=j) \quad (13)$$

Making use of equations (4), (12), and (13) allows the strategy to estimate $E(Y^l \mid S=m)$ for all combinations of *m* and *l*, and then to use these estimates to compute the different treatment effects γ_0^{ml}, α_0^{ml}, and θ_0^{ml}. Such an estimator is proposed in Section 3.

2.3.2 Composite effects

In the application a specific choice of weights will be considered, namely the unconditional distribution of treatments other than *m* in the population.

$$\tilde{v}^{m,l} = P(S = l \mid S \neq m), \qquad P(S = l \mid S \neq m) = \frac{P(S = l)}{1 - P(S = m)},\ m \neq l. \tag{14}$$

From a practical point of view an interesting question here is the following: Suppose we aggregate all observations not observed in treatment *m* in one group denoted by $-m$ without taking into account that this group is composed of different subgroups. Does an otherwise correctly performed estimation, that can be easily computed since it is based on the binary treatment model, correspond to a particular weighting scheme, and thus have a causal interpretation? The answer is yes, it has a causal interpretation, but that interpretation and thus the implied weighting scheme is different for $\gamma_0^m(v^m)$ and $\theta_0^m(v^m)$. Furthermore, it is difficult to derive the weights, denoted as $\breve{v}^m$ and $\breve{\breve{v}}^m$ in the following, explicitly, because they depend on the particular distribution of $P^m(X)$ in the specific comparison groups. This can be seen by the following considerations:

$$\gamma_0^m(\breve{v}^m) = EY^m - EY^{-m} =$$

$$= EY^m - E(Y^{-m} \mid S = -m)P(S = -m) - E(Y^{-m} \mid S = m)P(S = m)$$

$$= EY^m - E(Y^{-m} \mid S = -m)P(S = -m) - E(Y^{-m} \mid S = m)P(S = m)$$

$$= EY^m - \sum_{j=0}^{M} E(\breve{v}^{m,l}Y^l \mid S = l)P(S = l) - E_X\{E[\breve{v}^{m,l}Y^l \mid P^m(X), S = l] \mid S = m)\}P(S = m);$$

$$\theta_0^m(\breve{\breve{v}}^m) = E(Y^m \mid S = m) - E(Y^{-m} \mid S = m) =$$

$$= E(Y^m \mid S = m) - E_X\{E[Y^{-m} \mid P^m(X), S = -m] \mid S = m\}.$$

Whether this may or may not be a more sensible specification of the weights, depends on the context. It is however important to notice that $\theta_0^m(\tilde{v}^m)$ and $\theta_0^m(\breve{v}^m)$ are in general different causal effects.

3 A matching estimator

Given the choice probabilities, or a consistent estimate of them, the terms appearing in equations (12) and (13) can be estimated by any parametric, semiparametric, or nonparametric regression method that can handle one or two-dimensional explanatory variables. Lechner (1999b) proposes a matching estimator that is analogous to the rather simple algorithms used in the literature on binary treatments. It is given in Table 1.

Table 1: A matching protocol for the estimation of $\gamma_0^{m,l}$, $\alpha_0^{m,l}$, *and* $\theta_0^{m,l}$

Step 1	Specify and estimate a multinomial choice model to obtain $[\hat{P}_N^0(x), \hat{P}_N^1(x), ..., \hat{P}_N^M(x)]$.
Step 2	Estimate the expectations of the outcome variables conditional on the respective balancing scores. For a given value of *m* and *l* the following steps are performed: a) Compute $\hat{P}_N^{l\|ml}(x) = \frac{\hat{P}_N^l(x)}{\hat{P}_N^l(x) + \hat{P}_N^m(x)}$ or use $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$ directly. Alternatively, Step 1 may be omitted and the conditional probabilities may be directly modelled (as in the binary case; $\tilde{P}_N^{l\|ml}(x)$). b) Choose one observation in the subsample defined by participation in *m* and delete it from that pool. c) Find an observation in the subsample of participants in *l* that is as close as possible to the one chosen in step a) in terms of $\hat{P}_N^{l\|ml}(x)$, $\tilde{P}_N^{l\|ml}(x)$ or $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$. In case of using $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$ 'closeness' is based on the Mahalanobis distance. Do not remove that observation, so that it can be used again. d) Repeat a) and b) until no participant in *m* is left. e) 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.
Step 3	Repeat Step 2 for all combinations of *m* and *l*.
Step 4	Compute the estimate of the treatment effects using the results of Step 3.

Note: Lechner (1999b) suggests an estimator of the asymptotic standard errors for $\hat{\gamma}_N^{ml}$, $\hat{\alpha}_N^{ml}$, and $\hat{\theta}_N^{ml}$ based on the approximation that the estimation of the probabilities in Step 1 can be ignored.

Note that matching is done allowing the same comparison observation to be used repeatedly. This modification is necessary for the 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, this will always be the case when the number of participants is not equal in all treatments. This procedure has the potential problem that very 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 occurrence of this feature should be

checked, and if it appears, the algorithm needs to be suitably revised. Similar checks need to be performed – as usual – to make sure that the distributions of the balancing scores do indeed overlap sufficiently in the respective subsamples. For subsamples *m* and *l* this means that the distributions of $\hat{P}_N^{l|ml}(x)$ (or $\tilde{P}_N^{l|ml}(x)$ or $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$) have similar support.

The main advantage of the matching algorithm outlined in Table 1 is its simplicity. It is however not asymptotically efficient, since the issue of the typical trade-off in non-parametric regression between bias and variance is not addressed. Other more sophisticated and more computer intensive matching methods are discussed for example by Heckman, Ichimura, and Todd (1998).

The composite effects are estimated as aggregates of the pair-wise effect using the weights given in equation (14). In addition to that $\theta_0^m(\bar{v}^m)$ has been estimated directly from a pair-wise comparison in an aggregated sample using a probit model to estimate the respective probabilities and an accordingly simplified version of the algorithm outlined in Table 1.

4 Empirical application

4.1 Introduction and descriptive statistics

After experiencing increasing rates of unemployment in the mid 1990'ties Switzerland conducted a substantial active labour market policy. That policy has many different sub-programmes. For the purpose of this study they are aggregated into five different groups that contain more or less similar sub-programmes, i.e. NO PARTICIPATION in any programme, BASIC TRAINING (including courses of the local language and job counselling), FURTHER vocational TRAINING (including longer information technology courses as the largest part), EMPLOYMENT PROGRAMMES, and TEMPORARY subsidised EMPLOYMENT (job with company, labour office pays difference between wage and 70-80% of previous earnings[7]).

This application concentrates only on the largest Swiss canton, namely the canton of Zurich.[8] The population of interest is unemployed at the 31st of December 1997 (unemployment was a condition to be eligible for the programmes), aged between 25 and 55, has not participated in a

[7] This is slightly more than the unemployment benefits. Furthermore, the expiry date of unemployment benefits may be prolonged.

[8] Switzerland is divided into 26 cantons that enjoy a considerable autonomy from the central government.

substantial programme before the end of 1997, and is not disabled. The individual programme participation begins during 1998 and the observation period ends in March 1999. Hence, only short-run effects will possibly be discovered.

The data come from the Swiss unemployment registers and cover - before sample selection - the total population unemployed at that time in the canton of Zurich. Further information about the data base can be found in Gerfin and Lechner (1999).[9] The data base is fairly informative because it contains all the information the local labour offices use for the payment of the unemployment benefits and for advising the unemployed. Therefore, the conditional independence assumption is assumed to be valid for the remainder of this paper.[10]

Table 2: Descriptive statistics of selected variables according to the different states

		Non-participation	basic training	further training	employment programme	temporary employment
		median in subsample				
Age		39	38	40	40	39
Days of unemployment before start		251	218	219	335	247
Duration of programme in days		0	63	41	155	113
Starting day of programme after 1997		89	82	76	156	107
		share in subsample in %				
Gender: female		46	56	43	37	43
Subjective valuations of labour office						
Qualification:	best	57	42	79	51	60
	medium	19	22	11	24	19
	worst	24	35	10	24	21
Chance to find new job:	unclear	8	8	6	5	9
	very easy	2	1	3	1	2
	easy	11	9	16	9	15
	medium	55	55	62	59	58
	difficult	19	25	12	22	15
	special case	4	2	2	4	1
Native language: German		48	27	73	46	51
other than German, French, Italian		40	60	20	44	37
Number of observations		2822	1958	724	701	1463

Note: Starting dates for the nonparticipants are random draws in the distribution of all observable starting dates. Non-participants no longer unemployed at their designated starting date have been deleted from the sample.

[9] Gerfin and Lechner (1999) study the effects of the various programmes of the Swiss active labour market policy. Their data base covers all of Switzerland and also has some additional information from the pension system. Also, they consider more details of this policy. However, that data set is too expensive to handle for the current analysis.

[10] Obviously, there may be substantial arguments claiming that this may not be true. However, the aim of this study is to give an example of how an evaluation could be performed in this setting, and not to derive policy relevant conclusion. For the same reason, the reader is referred to Gerfin and Lechner (1999) for more discussion about the features of the single programmes.

Table 2 gives some descriptive statistics of selected variables for subsamples defined by the five different states. From these statistics it is obvious that the programmes are heterogeneous with respect to programme characteristics - for example duration - as well as with respect to individual characteristics of participants such as skills, qualifications, employment histories, among others.[11]

The effect of the programmes will be measured in terms of changes in the average probabilities of employment in the first labour market caused by the programme. It will be measured after the programme begins. The time in the programme is not considered as regular employment. This means that if somebody leaves a programme early in order to take up a job, this will influence our measure of effectiveness of the programme in a positive way. Such a measure could be disputed if one believes that being in the programme is a 'good thing' per se, but it is the approach taken in this paper to concentrate solely on the success in the labour market.

The entries in the main diagonal of Table 3 show the level of employment rates of the five groups in percentage points. The off-diagonal entries refer to the unadjusted difference of the corresponding levels. These rates are observed on a daily basis. The results in the table refer to latest observations available, i.e. to the end of March 1999. The last two columns refer to a composite category aggregating all states except the one given in the respective row.

Table 3: Unadjusted differences and levels of employment

	Non-participation	basic training	further training	employment programme	temporary employment	all other categories *	
Nonparticipation	(38.8)	8.6	-10.2	13.0	-9.7	0.9	(37.9)
basic training		(30.2)	-18.8	4.4	-18.3	-10.8	(41.0)
further training			(49.0)	23.2	0.5	11.9	(37.1)
employment programme				(25.8)	-22.7	-13.7	(38.3)
temporary employment					(48.5)	12.7	(35.8)

Note: The outcome variable is *employment* in %-points for day 451. Absolute levels on main diagonal and in the last column (shaded, in brackets). *All other categories* denotes the aggregation of all categories except the one given in the respective row.

The results show a wide range for average employment rates. The highest values that are close to 50% correspond to FURTHER TRAINING and TEMPORARY EMPLOYMENT. Clearly, the participants

[11] Unemployment duration until the beginning of training is an important variable for the participation decision. Since that variable is not observed for the group without treatment, starting dates are randomly allocated to these individuals according to the distribution of observed starting dates. Individuals no longer unemployed at the allocated starting dates are deleted from the sample. This approach follows closely an approach called *random* in Lechner (1999a).

with the worst (unadjusted) employment experience are the participants in EMPLOYMENT PROGRAMMES, followed by participants in BASIC TRAINING. However, from this table it is impossible to conclude whether the resulting order of employment rates is due to different impacts of the programmes, or to a selection of unemployed with already fairly different employment chances into to the various programmes. Disentangling these effects is of course the main task of every evaluation study.

4.2 Participation probabilities

Table 4 shows the estimation results of a multinomial probit model (MNP) using simulated maximum likelihood with the GHK simulator.[12] The largest group (non-participation) is chosen as the reference category. Although being fully parametric, the MNP is a flexible version of a discrete choice model, because it does not require the **I**ndependence of **I**rrelevant **A**lternatives assumption to hold.[13]

The variables that are used in the MNP are selected by a preliminary specification search based on binary probits (each relative to the reference category) and score tests against omitted variables. Based on that step the final specifications contain a varying number of mainly discrete variables that cover groups of attributes related to personal characteristics, valuations of individual skills and chances on the labour market as assessed by the labour office, previous and desired future occupations, as well as information related to the previous unemployment spell and to the unemployment spell that was still on-going in the last day of 1997. Entries for variables excluded from a particular choice equation show a 0 for the coefficient and '-' for the standard error.

[12] See for example Börsch-Supan, Hajivassiliou (1993) and Geweke, Keane, Runkle (1994).

[13] In practise, some restrictions on the covariance matrix of the errors terms of the MNP need to be imposed, because not all elements of the covariance matrix are identified and to avoid excessive numerical instability. See below.

Table 4: Results of the estimation of a multinomial probit model

	basic training		further training		employment programme		temporary employment	
	coef.	std.	coef.	std.	coef.	std.	coef.	std.
Constant	**.97**	.29	-2.55	1.7	-2.19	1.6	-.54	.56
Age in years / 10	.01	.02	.14	.10	*.22*	.10	*.10*	.05
Older than 50 years	0	-	0	-	0	-	-.15	.12
Gender: female	**.24**	.07	*-.53*	.27	*-.39*	.22	0	-
Married	0	-	0	-	*-.44*	.22	0	-
First foreign language: English	.06	.07	0	-	0	-	0	-
French, Italian, German	**-.22**	.07	0	-	*.42*	.22	0	-
Native language: French	**.54**	.18	0	-	0	-	0	-
Italian	**.52**	.14	*-1.00*	.54	0	-	0	-
other than French, Italian, German	**.56**	.15	*-.89*	.46	0	-	-.15	.10
Permanent foreign resident (work permit C)	0	-	0	-	-.55	.27	0	-
Temporary foreign resident (work permit B)	**.27**	.08	*-1.05*	.58	-.25	.23	0	-
Information about local labour office								
located in labour market region: Small villages	**-1.49**	.37	-.38	.66	-.94	.62	-.07	.34
located in labour market region: Big cities	**-.63**	.13	*-.95*	.47	*-.79*	.36	-.17	.14
share of entry into long-term unemployed of all UE	**-4.86**	1.9	4.31	5.4	9.32	5.8	-.76	2.3
no information on shares available	**-1.23**	.32	.14	.70	.36	.54	.00	.32
Subjective valuations of labour office								
qualification: best [a]	**-.16**	.06	.41	.27	-.12	.12	0	-
worst [a]	0	-	0	-	0	-	0	-
chance to find a new job: unclear [b]	0	-	-.56	.37	*-.62*	.33	.07	.11
very easy [b]	0	-	0	-	0	-	.18	.20
easy [b]	0	-	0	-	0	-	*.23*	.11
difficult [b]	*.11*	.07	*-.71*	.37	0	-	*-.22*	.11
special case [b]	**-.37**	.15	-.81	.53	-.13	.26	**-1.35**	.49
Desired level of occupation: part time [c]	**-.27**	.08	0	-	**-.70**	.30	**-.37**	.13
full time [c]	*.09*	.05	0	-	-.02	.09	-.06	.07
Desired occupation same as last occupation [d]	**-.11**	.05	0	-	0	-	0	-
Last sector								
agriculture	*-.50*	.22	-1.48	1.3	0	-	0	-
construction	-.06	.07	-.41	.30	-.35	.22	.02	.11
public services	**.25**	.09	0	-	0	-	*-.33*	.19
communications, news	.50	.31	1.00	.92	0	-	.53	.47
repairs	*.35*	.16	0	-	0	-	0	-
tourism, catering	0	-	*-.89*	.47	0	-	0	-
services (properties, renting, leasing, ...)	0	-	0	-	*-1.70*	.96	0	-
other services	0	-	0	-	0	-	**.45**	.19
sectoral unemployment rate in % * 10	0	-	0	-	-.28	.17	0	-

Table 4 to be continued.

Table 4 - continued: Results of the estimation of a multinomial probit model

	basic training		further training		employment programme		temporary employment	
	coef.	std.	coef.	std.	coef.	std.	coef.	std.
Last occupation								
construction	0	-	-1.04	.64	0	-	*.35*	.18
transportation	*-.25*	.12	0	-	0	-	0	-
metals	0	-	0	-	0	-	0	-
painting, drawing, ...	0	-	0	-	0	-	*.34*	.16
office	0	-	*.70*	.36	0	-	0	-
tourism, catering	*.13*	.06	0	-	0	-	0	-
management, judicial system, self-employed, ...	0	-	0	-	-.36	.32	0	-
architects, engineers, technicians	0	-	*1.45*	.77	0	-	0	-
security, social services, ...	0	-	-.88	.54	0	-	0	-
cosmetics, and similar services	0	-	-.88	.54	0	-	*1.02*	.49
education	0	-	0	-	0	-	**.81**	.31
Previous job position: high (management, ...)	0	-	0	-	*-.80*	.41	*-.26*	.14
very low	0	-	*-.81*	.41	0	-	0	-
Previous monthly earnings: below 2000 SFr	0	-	.43	.27	0	-	0	-
above 5000 SFr	0	-	.06	.19	0	-	*-.18*	.09
Duration of previous unemployment spell / 1000	**-.25**	.06	0	-	*.90*	.49	0	-
Duration of CUES until start of programme / 1000	.97	.77	-.58	.77	-1.09	.73	**-2.10**	.68
(Duration of CUES until 31/12/97 / 1000)^2 * 10	**-.79**	.23	0	-	0	-	0	-
Duration of CUES less than 90 days	0	-	-.40	.26	*-.93*	.46	0	-
less than 180 days	0	-	0	-	-.70	.33	*-.27*	.12
less than 270 days	0	-	0	-	-.31	.21	0	-
Participation in programme of < 2 weeks in CUES	0	-	.54	.40	0	-	0	-
Days from 12/31/97 until start / 100	*-.16*	.09	-.08	.16	**.56**	.22	**.36**	.12

Implied covariance matrix of the error terms*

	coef.	t-val.	coef.	t-val.	coef.	t-val.	coef.	t-val.	coef.	t-val.
Nonparticipation	1	-	0	-	0	-	-.23	-.3	-.02	-.05
Basic training			0.38	1.0	-1.71	-1.7	-.55	-1.5	-.03	-.2
Further training					8.76	-	1.54	-1.1	-.22	-.6
Employment programme							2.78	-	-.51	-1.4
Temporary employment							-		1.96	-

Implied correlation matrix of the error terms

Nonparticipation	1	0	0	-.14	-.02
Basic training		1	-.94	-.53	-.04
Further training			1	.31	-.05
Employment programme				1	-.22

Note: Simulated maximum likelihood estimates using the GHK simulator (140 draws in simulator for each observation and choice equation). Coefficients of the category NONPARTICIPATION are normalized to zero. Inference is based on the outer product of the gradient estimate of the covariance matrix of the coefficients ignoring simulation error.
N = 7669. Value of log-likelihood function: -10188.73.
Bold numbers indicate significance at the 1% level (2-sided test), numbers in *italics* relate to the 5% level.
If not stated otherwise, all information in the variables relates to the last day in December 1997. a) Reference group: qualification: medium; b) Reference group: chance to find a new job: medium; c) Reference group: unknown desired level (about 35% of sample); d) Based on the 3 digit job classification.
*) 9 Cholesky factors are estimated to ensure that the covariance of the errors remains positive definite. t-values refer to the test whether the corresponding Cholesky factor is zero (off-diagonal) or one (main-diagonal).

The estimation results show that compared to the status NONPARTICIPATION the coefficients related to the choice equations are fairly heterogeneous, including sign changes of significant variables. Although this could be expected already from the descriptive statistics given in Table 2, the MNP again confirms this finding and also shows that it is related to more variables than those given in Table 2. All these results basically confirm the view that individuals with severe problems on the labour market have a clearly higher probability of ending up in either BASIC TRAINING or an EMPLOYMENT PROGRAMME. The latter is particularly used as a programme for the long-term unemployed.[14] In contrast it is more likely for the 'easier' cases to participate IN FURTHER TRAINING or TEMPORARY EMPLOYMENT. Therefore, one can safely conclude that the various groups of active labour market policies are targeted to different groups of the unemployed.

The lowest part of Table 4 gives the estimated covariance matrix of the error terms as well as the implied correlation matrix. The estimated standard errors of the error terms vary between .6 (BASIC TRAINING) and about 3 (FURTHER TRAINING). The estimated correlations are between -.9 and .3. The high negative correlation as well as the general lack of precision of the covariance matrix estimate is a somewhat worrying feature.[15] The lack of precision is transferred to the other estimated coefficients. Compared to a more restrictive specification, there appears to be a considerable increase in the standard errors of the two groups with the largest estimated variances, namely FURTHER TRAINING and EMPLOYMENT PROGRAMME.

The estimation results presented in Table 4 are used to compute the participation probabilities of the various categories conditional on *X*. Table 5 gives some descriptive statistics of the distribution of these probabilities in the various subgroups. The columns of the upper part of that table contain the 5%, 50%, and 95% quantiles of the distribution of the respective probabilities as they appear in the sample denoted in the particular row. Of course, the values of the probabilities that correspond to the probabilities of the category in which these observations are observed (shaded area) are the highest one in each column. Another observation is that there is

[14] Note that for EMPLOYMENT PROGRAMMES the reference group of the dummy variables measuring length of the current unemployment spells is more than 270 days. Combining the coefficients for the dummy variables with the coefficient of the continuous variable gives, for example, a value of -2 for 50 days of unemployment compared to a value of -.3 for 300 days of unemployment. Ignoring the continuous variable, that is insignificant, gives corresponding values of –1.9 and 0. In addition, the begin of the programme, that is also positively related to the duration of the unemployment spell, is significantly later for participants in EMPLOYMENT PROGRAMMES than for other programmes.

[15] Increasing the number of draws from 140 to 250 gives basically the same result.

considerable variation of the probabilities. This means on the one hand that the observations within a treatment show a considerably heterogeneity with respect to their characteristics. But on the other hand, there is probably sufficient overlap as is necessary for the successful working of the matching (and of course every other nonparametric) procedure.[16]

Table 5: Descriptive statistics for the distribution of the participation probabilities computed from the multinomial probit in the population and the subsamples

	Quantiles of probabilities in %											
	basic training			further training			employment programme			temporary employment		
Samples	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
Nonparticipation	5	21	49	1	8	23	1	6	25	8	18	32
Basic training	10	33	71	1	4	22	.4	5	23	5	15	29
Further training	5	19	42	3	15	27	1	5	23	9	18	33
Employment programme	4	18	45	1	6	21	3	15	36	9	19	33
Temporary employment	4	19	48	1	8	23	1	7	27	11	21	38
All	5	23	57	1	7	23	1	6	26	7	18	33

	Correlation matrix of probabilities in full sample			
	basic training	further training	employment programme	temporary employment
Nonparticipation	-.49	.07	-.23	-.06
Basic training		-.46	-.36	-.57
Further training			-.20	.09
Employment programme				.14

Note: See note below Table 4.

The lower part of Table 5 presents the correlations of these probabilities in the complete sample. There are fairly strong negative correlations between the probabilities for some treatments, but they do not get smaller than -0.6 for any pair. Although, the magnitudes of these correlations change somewhat for the subsamples defined by treatment status, they have a very similar structure in these subsamples (not given here).

There are three additional probabilities that are used subsequently in the matching. First, for the estimator that will be called *naive conditional*, a probit model based only on observations observed in group *m* and *l* is used to obtain $\tilde{P}^{m|ml}(X)$. The respective explanatory variables are those that influence the choices between both *m* and *l* and the reference category

[16] Note that matching as implemented here is with replacement. Therefore, it is less demanding in terms of distributional overlap than matching without replacement, because extreme observations in the comparison group, that are the rare commodity in that trade, can be used more than once.

(NONPARTICIPATION). Secondly, $P^{m|ml}(X)$ is computed using the definition of conditional probabilities given exclusive categories $P^{m|ml}(X) = P^m(X)/[P^m(X) + P^l(X)]$, where $P^m(X)$ and $P^l(X)$ are the probabilities from the multivariate probit model (*MVP conditional*). If the MNP is the correct specification, then the probabilities $\tilde{P}^{m|ml}(X)$ are misspecified, because they ignore the dependence of $P^{m|ml}(X)$ on variables influencing other choices and also use another functional form than $P^{m|ml}(X)$. Finally, $P^{-m}(X)$ is estimated with a probit model using all the explanatory variables appearing in Table 4.

4.3 Matching using different balancing scores

4.3.1 Quality of the matches

Matching is implemented as described in Table 1. The Mahalanobis distance is used as distance metric when matching is on more than one variable, i.e. $P^m(X)$ and $P^l(X)$ (called *MVP unconditional* in the following). This metric has some appeal because the probabilities are fairly close to being continuous.

Using the standardised bias as indicator of the match-quality, the results given in Table 6 show that match quality is rather good with respect to the probabilities used for the matching. This indicates that the overlap of these probabilities is generally sufficient. An exception is perhaps the case when the small and fairly special group of individuals participating in EMPLOYMENT PROGRAMMES and FURTHER TRAINING are used as a comparison groups. This problem appears however mainly in the case of using both $P^m(X)$ and $P^l(X)$, and to a much lesser extend for matching on single probabilities.

Table 6: Are the probabilities used for matching balanced ? Results for the absolute standardised bias (100)*

l / *m*	Nonparticipation		basic training		further training		employment programme		temporary employment	
	MVP unconditional, $P^m(X)$ and $P^l(X)$									
	$P^m(X)$	$P^1(X)$	$P^m(X)$	$P^2(X)$	$P^m(X)$	$P^3(X)$	$P^m(X)$	$P^4(X)$	$P^m(X)$	$P^5(X)$
Nonparticipation	0	0	.5	.8	.8	1.3	1.2	2.9	.1	.6
Basic training	.6	.4	0	0	4.0	2.5	3.9	1.6	1.0	1.5
Further training	.3	.4	.6	.6	0	0	2.0	4.8	.7	.9
Employment programme	.3	.1	.5	.7	1.8	3.0	0	0	.9	1.0
Temporary employment	.4	.5	.5	.6	.6	1.4	1.0	1.2	0	0
	MVP conditional ($P^{m\|ml}$)									
	$P^{m\|ml}$		$P^{m\|ml}$		$P^{m\|ml}$		$P^{m\|ml}$		$P^{m\|ml}$	
Nonparticipation	0		.0		.3		.3		.0	
Basic training	.1		0		.2		.2		.1	
Further training	.1		.0		0		1.6		.1	
Employment programme	.0		.1		.2		0		.0	
Temporary employment	.2		.0		.3		.4		0	
	Naive conditional ($\tilde{P}^{m\|ml}$)									
	$\tilde{P}^{m\|ml}$		$\tilde{P}^{m\|ml}$		$\tilde{P}^{m\|ml}$		$\tilde{P}^{m\|ml}$		$\tilde{P}^{m\|ml}$	
Nonparticipation	0		.0		.3		.2		.0	
Basic training	.1		0		.4		.4		.2	
Further training	.1		.0		0		2.1		.1	
Employment programme	.0		.2		.5		0		.1	
Temporary employment	.2		.1		.4		.1		0	
	MVP unconditional, $P^m(X)$									
	$P^m(X)$		$P^m(X)$		$P^m(X)$		$P^m(X)$		$P^m(X)$	
Nonparticipation	0		.0		.3		.1		.0	
Basic training	.3		0		3.3		1.0		.6	
Further training	.1		0		0		.8		.1	
Employment programme	.1		.1		.3		0		.1	
Temporary employment	.1		.1		.1		.8		0	

Note: The absolute standardized bias (SB) is defined as the absolute difference of the means in the subsamples *m* and the matched comparison sample obtained from participants in *l* divided by the square root of the average of the variances in *m* and the matched comparison sample * *100*. SB* can be interpreted as bias in % of the average standard deviation.

However, the real question is whether matching on these probabilities is sufficient to balance the covariates. Table 7 gives the results of two summary measures - the median absolute standardised bias and the mean squared standardised bias - that give an indication of the distance between the marginal distributions of the covariates that influence the choice in group *m* and the matched comparison group *l*. There does not appear to be a consensus in the literature about how to measure the distance between high dimensional multivariate distributions, but the two measures given are often used. Their major shortcoming is that they are based on the (weighted)

differences of the marginal means only, thus ignoring any other feature of the respective multivariate distributions.

Table 7: Are the covariates balanced ? Results for the median absolute standardised bias (MASB) and the mean squared standardised bias (MSSB)

l / *m*	Nonparticipation		basic training		further training		employment programme		temporary employment	
	MASB	MSSB	MASB	MSSB	MASB	MSSB	MASB	MSSB	MASB	MSSB
	MVP unconditional, $P^m(X)$ and $P^l(X)$									
Nonparticipation	0	0	3	16	4	31	6	56	2	12
Basic training	3	18	0	0	6	111	9	136	3	21
Further training	4	25	3	28	0	0	8	108	4	33
Employment p.	4	27	4	N/A	5	61	0	0	4	33
Temporary e.	3	16	4	21	3	26	6	70	0	0
Sum	*14*	*86*	*14*	*N/A*	*18*	*229*	*29*	*370*	*13*	*99*
	MVP conditional, $P^{m\|ml}$									
Nonparticipation	0	0	3	13	5	45	7	120	3	17
Basic training	3	22	0	0	6	134	8	209	3	28
Further training	3	13	4	31	0	0	9	248	4	33
Employment p.	3	21	5	N/A	6	76	0	0	3	31
Temporary e.	2	15	3	22	5	38	6	112	0	0
Sum	*11*	*71*	*15*	*N/A*	*22*	*293*	*30*	*689*	*13*	*109*
	Naive conditional, $\tilde{P}^{m\|ml}$									
Nonparticipation	0	0	3	18	5	54	7	128	2	13
Basic training	3	19	0	0	6	141	8	221	2	21
Further training	2	23	3	23	0	0	9	287	3	34
Employment p.	4	29	4	29	6	N/A	0	0	3	38
Temporary e.	3	17	4	26	4	39	6	79	0	0
Sum	*12*	*88*	*14*	*96*	*21*	*N/A*	*30*	*715*	*10*	*106*
	MVP unconditional, $P^m(X)$									
Nonparticipation	0	0	7	90	11	355	7	365	4	44
Basic training	3	19	0	0	11	426	10	402	4	36
Further training	4	32	5	61	0	0	7	377	5	47
Employment p.	4	37	9	137	9	370	0	0	5	53
Temporary e.	3	17	7	92	14	413	8	346	0	0
Sum	*14*	*105*	*28*	*380*	*45*	*1564*	*32*	*1490*	*18*	*180*

Note: The standardized bias (SB) is defined as the difference of the means in the respective subsamples divided by the square root of the average of the variances in *m* and the matched comparison sample obtained from participants in *l* * 100. *SB* can be interpreted as bias in % of the average standard deviation. The median of the absolute standardized bias (MASB) and the mean of the squares of the standardized bias (MSSB) are taken with respect to all 56 covariates included in the estimation of the MVP (see Table 4). N/A: Not available, because one covariate has zero variance for that pair.

Using the results in Table 7 to rank the different versions according to their match quality, the first conclusion is that matching solely on the marginal probabilities gives a comparatively bad

match for most of the combinations of different treatments. Thus, the theoretical finding that conditioning on the marginal probability is not sufficient seems to matter in this application.

Comparing the match quality obtained by using the two different ways to estimate the conditional probabilities, it is very hard to spot any systematic difference.

Counting the cells where one estimator dominates the other, the estimator matching on both marginal probabilities appears to be superior to all the others. Compared to the version using only the conditional probability computed from the MNP, it dominates slightly on MASB (+3) and considerably on MSSB (+ 8). With respect to the naive estimator, the ranking is similar. The naive conditional is clearly dominated for MSSB, whereas no difference appears for MASB. Since MSSB is more influenced by extreme values than MASB, it appears that the unconditional MVP produces less extreme mismatches than both of the other methods, but particularly than the naive conditional. Furthermore, the dominance is most visible for the control groups that appear most difficult to match according to Table 6, namely FURTHER TRAINING and EMPLOYMENT PROGRAMMES.

A matching algorithm that uses every control group only once runs into problems in regions of the attribute space where the density of the probabilities is very low for the control group compared to the treatment group.[17] An algorithm that allows to use the same observation more than once, does not have that problem as long as there is an overlap in the distributions. The drawback of that estimator might be that it uses observations too often, in the sense that comparable observations that are almost identical to the ones actually used are available. Hence, in principle there could be substantial losses in precision as a price to pay for a reduction of bias.

Table 8 addresses that issue by considering two measures. The first is a concentration ratio computed as the sum of weights in the first decile of the weight distribution – each weight equals the number of treated observations the specific control observation is matched to – divided by the total sum of weights in the comparison sample. The second measure gives the mean weights of the matched comparison observations. In case the comparison sample is smaller than the treatment sample so that the mean must be larger than 1, the mean is adjusted downwards by the ratio of the sample sizes. Note that it is not possible to attribute the numbers in Table 8 to an excess use of 'middle of the road' observations (which is not desired and could be avoided in

[17] Note that since every group acts as a comparison group in the multi-programme framework, this occurs by definition.

principle), or to a very thin density in a region with many treatment observations (which is unavoidable).

Table 8: Excess use of single observations

l / *m*	Nonparticipation		basic training		further training		employment programme		temporary employment	
	top10	mean	top10	mean	top10	mean	top10	mean	top10	mean
	MVP unconditional, $P^m(X)$ and $P^l(X)$									
Nonparticipation			31	1.7 a)	42	1.2 a)	45	1.3 a)	26	1.4 a)
Basic training	30	1.7			57	1.5 a)	51	1.6 a)	36	1.9 a)
Further training	20	1.2	24	1.5			45	2.7 a)	25	1.5
Employment programme	23	1.3	26	1.7	39	2.7			25	1.6
Temporary employment	24	1.4	29	1.9	39	1.5 a)	41	1.6 a)		
	MVP conditional, $P^{m\|ml}$									
Nonparticipation			31	1.7 a)	43	1.2 a)	48	1.3 a)	30	1.5 a)
Basic training	31	1.8			58	1.6 a)	51	1.8 a)	38	1.9 a)
Further training	21	1.3	26	1.7			46	2.9 a)	27	1.6
Employment programme	24	1.4	29	1.7	41	3.0			27	1.6
Temporary employment	24	1.5	29	2.0	38	1.6 a)	39	1.6 a)		
	Naive conditional, $\tilde{P}^{m\|ml}$									
Nonparticipation			30	1.8 a)	43	1.3 a)	47	1.4 a)	30	1.3 a)
Basic training	32	1.8			56	1.7 a)	53	1.8 a)	36	1.9 a)
Further training	21	1.3	25	1.6			46	2.8 a)	25	1.6
Employment programme	23	1.4	31	1.8	41	2.8			27	1.6
Temporary employment	24	1.5	31	1.9	42	1.6 a)	40	1.6 a)		
	MVP unconditional, $P^m(X)$									
Nonparticipation			29	1.3 a)	27	1.1 a)	30	1.2 a)	28	1.3 a)
Basic training	31	1.8			46	1.5 a)	43	1.5 a)	38	1.8 a)
Further training	20	1.3	24	1.6			32	2.3 a)	23	1.5
Employment programme	23	1.4	28	1.2	32	2.4			26	1.6
Temporary employment	24	1.5	28	1.9	29	1.4 a)	29	1.3 a)		

Note: *top10*: Share of the sum of largest 10% of weights of total sum of weights. *Mean*: Mean of positive weights, a) Mean adjusted (multiplied) by N^l/N^m, because N^m is larger than N^l.

A first conclusion from that table is that the higher number for the means and in particular the concentration ratios appear for exactly the treatments that already showed up in Table 7 as ones that made up the worst comparison groups, namely FURTHER TRAINING and EMPLOYMENT PROGRAMMES. These are also the programmes with the smallest number of observations. Whereas for the other treatments, no real differences appear across estimation methods, for those two, the estimator based on only one marginal probability uses considerably more observations than the other estimators. However, it appeared already in Table 7 that this results in biases (in

terms of balancing the distributions) that are considerably higher than those for the other estimators. With respect to the other estimators in most cases the MVP unconditional with two probabilities uses more observations than the other estimators. Although the differences are small, they are fairly systematic. This result is surprising because that estimator appeared also to be the best in terms of bias (Table 7). Hence it appears that – at least in this application - this estimator has favourable properties with respect to bias as well as with respect to precision.

4.3.2 The sensitivity of the evaluation results

In this section the issue whether the final evaluation results are sensitive with respect to the choice of the balancing scores is addressed. To avoid an excess of numbers, Table 9 gives the estimation results for the various pair-wise effects for $\theta_0^{m,l}$ only. A positive number indicates that the effect of the treatment shown in the row compared to the treatment denoted in the column is an additional amount of XX%-points of employment. This effect is valid for the population appearing in the rows of the table. For example, the entry for the fifth treatment in the row and the second treatment in the column (MVP unconditional, $P^m(X)$ and $P^l(X)$) should be read as 'for the population participating in TEMPORARY EMPLOYMENT, TEMPORARY EMPLOYMENT increases the probability of being employed on day 451 on average by 12.8 %-points compared to that population being in BASIC TRAINING'. In addition to the results obtained by the different estimation methods, the lower part of the table repeats the unadjusted difference to give an impression on how much the estimators correct this difference for potential selection bias due to observable differences in the different groups.

Table 9: Estimation results for $\theta_0^{m,l}$

l / m	Nonparticipation	basic training	further training	employment programme	temporary employment
	MVP unconditional, $P^m(X)$ and $P^l(X)$				
Nonparticipation		2.1 (2.0)	*-7.0* (3.5)	3.9 (3.4)	**-8.2 (2.1)**
Basic training	**-5.9** (2.1)		-5.6 (6.5)	1.9 (4.7)	**-12.5 (3.3)**
Further training	1.7 (2.9)	5.5 (3.0)		3.2 (5.1)	-2.8 (3.3)
Employment programme	-1.3 (2.9)	-0.6 (3.2)	-5.0 (4.9)		**-11.4 (3.3)**
Temporary employment	**7.5** (2.2)	**12.8** (2.5)	2.0 (3.7)	**13.9** (3.5)	
	MVP conditional, $P^{m\|ml}$				
Nonparticipation		*5.1* (2.1)	**-11.9** (4.3)	-2.3 (5.4)	**-10.5** (2.3)
Basic training	*-5.3* (2.4)		*-14.2* (6.9)	-7.7 (6.2)	**-14.6** (3.1)
Further training	0.0 (3.0)	*6.9* (3.3)		13.0 (7.5)	**-9.5** (3.4)
Employment programme	-5.6 (3.0)	-2.1 (3.4)	-7.3 (5.4)		**-13.7** (3.4)
Temporary employment	**5.9** (2.2)	**11.0** (2.5)	*9.2* (4.0)	**14.4** (4.1)	
	Naive conditional, $\tilde{P}^{m\|ml}$				
Nonparticipation		3.2 (2.1)	-2.7 (4.3)	-1.8 (5.4)	**-11.7** (2.3)
Basic training	-4.4 (2.4)		-7.2 (8.2)	-7.5 (6.6)	**-15.9** (3.1)
Further training	4.4 (2.9)	3.6 (3.2)		11.7 (8.2)	**-9.4** (3.4)
Employment programme	-2.7 (3.0)	-4.3 (3.5)	3.3 (5.7)		**-12.3** (3.3)
Temporary employment	**9.1** (2.2)	**11.9** (2.6)	*9.0* (4.4)	6.0 (4.3)	
	MVP unconditional, $P^m(X)$				
Nonparticipation		3.3 (2.0)	**-9.0** (2.6)	**8.9** (2.5)	**-11.6** (2.1)
Basic training	**-7.0** (2.5)		**-15.9** (5.5)	*10.3* (4.1)	**-12.9** (3.2)
Further training	5.0 (2.9)	5.0 (3.1)		**14.4** (3.9)	*-7.2* (3.2)
Employment programme	-0.7 (3.0)	4.6 (3.2)	-8.0 (4.4)		**-17.5** (3.4)
Temporary employment	**6.7** (2.2)	**13.5** (2.5)	-0.5 (3.1)	**22.0** (2.8)	
	Levels and unadjusted raw differences (same as in Table 3)				
Nonparticipation	(38.8)	8.6	-10.2	13.0	-9.7
Basic training	-8.6	(30.2)	-18.8	4.4	-18.3
Further training	10.2	18.8	(49.0)	23.2	0.5
Employment programme	-13.0	-4.4	-23.2	(25.8)	-22.7
Temporary employment	9.7	18.3	-0.5	22.7	(48.5)

Note: The outcome variable is *employed* for day 451 (in %-points). Standard errors are in brackets. **Bold** numbers indicate significance at the 1% level (2-sided test), numbers in *italics* indicate significance at the 5% level.

Comparing the different estimators it appears that the biases observed for MVP unconditional with only one probability, that appeared already in Table 7, lead also to biased results. The bias appears to be particularly severe for the groups where the match quality is worst, namely when FURTHER TRAINING and EMPLOYMENT PROGRAMMES act as the comparison programmes.

Comparing the other three estimators it appears first of all that the use of more comparison observations by MVP unconditional using two probabilities (MVPUC2) results – as expected –in

somewhat smaller standard errors, particularly so for the more difficult cases of having FURTHER TRAINING and EMPLOYMENT PROGRAMMES as comparison groups.

Comparing the results column by column it appears that we get fairly similar conclusions from the three estimators when NONPARTICIPATION and BASIC TRAINING are used as comparison states. The adjustment works in the same direction. For the case of comparisons to FURTHER TRAINING and EMPLOYMENT PROGRAMMES, MVPUC2 gives somewhat different results compared to the two other estimators. This is however expected, since it appears to match the distribution of attributes better in these cases (see Table 7). A puzzling effect occurs for the comparison of FURTHER TRAINING with TEMPORARY EMPLOYMENT. Although matching appears to have worked very similarly for all three estimators – this is confirmed by checking the matches variable by variable – the coefficient for MVUC2 is about two standard deviations apart from the coefficients of the other two estimators (and hence much closer to the unadjusted difference).[18]

4.4 Heterogeneity of the effects

In this section heterogeneity with respect to the results is considered in more detail. Since MVPUC2 performed best in terms of match quality the following results are based on MVPUC2 only. Let us first consider the heterogeneity with respect to the different programmes for a person randomly selected from the population, given in the upper part of Table 10.[19] It is obvious that the programmes have different impacts. BASIC TRAINING is the only programme that has negative effects compared to NONPARTICIPATION. It is also dominated by FURTHER TRAINING and TEMPORARY EMPLOYMENT. Similarly, EMPLOYMENT PROGRAMMES are dominated by FURTHER TRAINING and TEMPORARY EMPLOYMENT. TEMPORARY EMPLOYMENT dominates all other programmes, with the exception of FURTHER TRAINING (but see the discussion in the previous section). All other effects are not significant.

It appears to be surprising that although the various groups of participants in the different programmes are very heterogeneous, the effects for these different populations are not. This can be seen by comparing the corresponding numbers above and below the diagonal in the lower part of the table (i.e. comparing $\theta_0^{m,l}$ to $\theta_0^{l,m}$). Such a finding could suggest that these programmes are not well targeted, in the sense that a person participated in an EMPLOYMENT PROGRAMME

[18] However, the estimated values for $\gamma_0^{m,l}$ are again very close for all three estimators.

although having a higher expected employment probability in TEMPORARY EMPLOYMENT for example. We obtain similar conclusions by comparing $\gamma_0^{m,l}$ to $\alpha_0^{m,l}$ and to $\theta_0^{m,l}$. The overall conclusion from that is that treatment heterogeneity is important, but population heterogeneity with respect to the effects is not.

Table 10: Estimation results for $\gamma_0^{m,l}$, $\alpha_0^{m,l}$, and $\theta_0^{m,l}$ (MVP unconditional)

l / *m*	Nonparticipation	basic training	further training	employment programme	temporary employment
	$\gamma_0^{m,l} = E(Y^m - Y^l)$				
Nonparticipation		*3.7* (1.8)			**-7.4** (2.0)
Basic training	*-3.7* (1.8)		*-7.8* (3.5)		**-11.2** (2.2)
Further training		*7.8* (3.5)		*8.9* (4.3)	
Employment programme			*-8.9* (4.3)		**-12.3** (3.3)
Temporary employment	**7.4 (2.0)**	**11.2 (2.2)**		**12.3** (3.3)	
	$\alpha_0^{m,l} = E(Y^m - Y^l \mid S = m,l)$				
Nonparticipation		*3.7* (1.8)	-5.9 (3.1)		**-7.9** (1.9)
Basic training	*-3.7* (1.8)				**-12.6** (2.5)
Further training	5.9 (3.1)				
Employment programme					**-13.1** (3.0)
Temporary employment	**7.9 (1.9)**	**12.6** (2.5)		**13.1** (3.0)	
	$\theta_0^{m,l} = E(Y^m - Y^l \mid S = m)$				
Nonparticipation			*-7.0* (3.5)		**-8.2** (2.1)
Basic training	**-5.9** (2.1)				**-12.5** (3.3)
Further training		5.5 (3.0)			
Employment programme					**-11.4** (3.3)
Temporary employment	**7.5** (2.2)	**12.8** (2.5)		**13.9** (3.5)	

Note: The outcome variable is *employed* for day 451 (in %-points). Standard errors are in brackets. **Bold** numbers indicate significance at the 1% level (2-sided test), numbers in *italics* indicate significance at the 5% level. Estimated coefficients not significant at the 10% level are omitted. All estimates are based on a balancing score including both marginal probabilities.

The following figure gives an idea about the dynamics of the effects. Furthermore, questions whether the effects are homogenous with respect to groups that are considered to have labour market problems can be addressed as well. Figure 1 presents the effect of each programme compared to nonparticipation for the total population ($\gamma_0^{m,l}$).[20] Effects are displayed starting in

[19] The term population refers to the population defined by selection rules explained before.

[20] $\gamma_0^{m,1}$ is a better measure than $\theta_0^{m,1}$ for a benchmark for the overall performance of the programmes compared to NONPARTICIPATION, because its reference population is independent of the specific programme.

mid 1998 up to March 1999. A value larger than zero indicates that NONPARTICIPATION would actually increase employment shares compared to the specific programme.

Figure 1: Effects of NONPARTICIPATION compared to the programmes for the population ($\gamma_0^{NP,m}$ for employment)

Figure 1.1: All

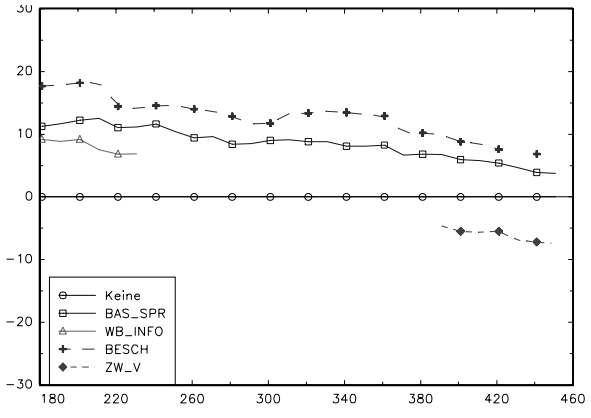


Figure 1.2: Unemployed for more than 270 days before the programme

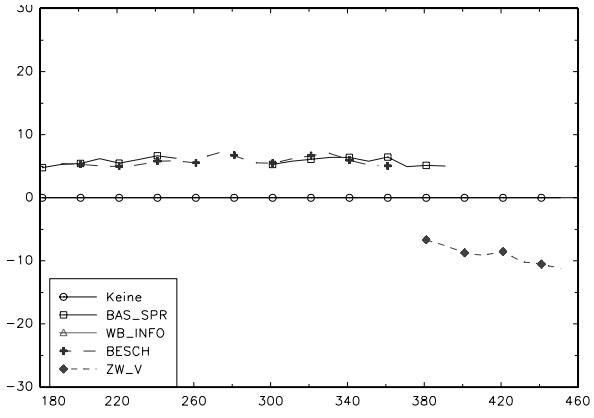


Figure 1.3: Native language not a Swiss language

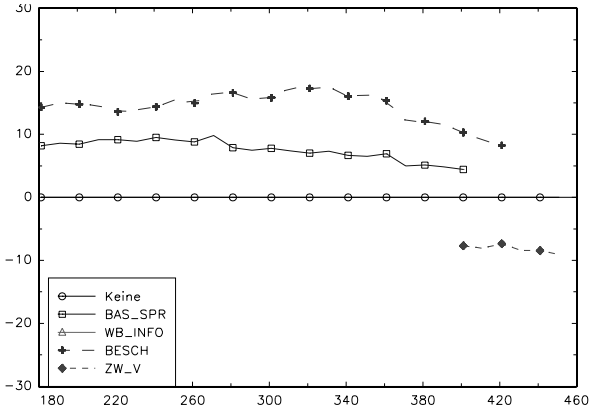
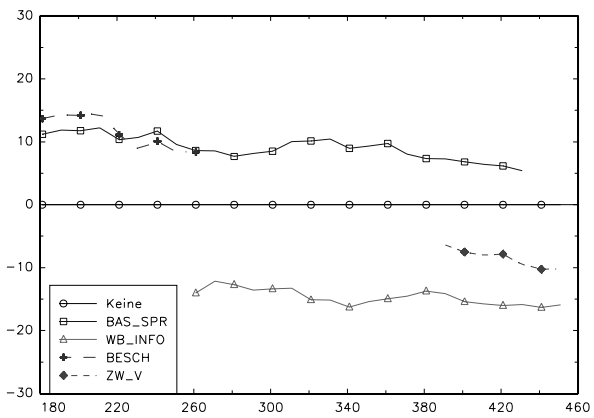


Figure 1.4: Women



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

It appears that only TEMPORARY EMPLOYMENT has positive effects. However, the dynamics suggest that time might work in favour of the programmes and could lead to different long-term results. Similar results appear for people whose unemployment spell before the programme already exceeded 3 quarters (270 days) as well as for individuals whose native language is not

German, French, or Italian. A striking difference shows up for another group considered to have special problems on the labour market, namely women. Here, FURTHER TRAINING appeared to have stable large positive effects in the magnitude of about 15 %-points.

Figure 2: Effects of NONPARTICIPATION compared to the programmes for the population ($\gamma_0^{NP,m}$ for employment): Time relative to start of programme

Figure 2.1: All

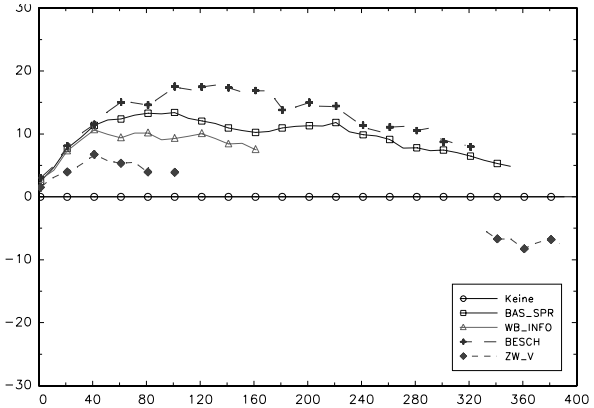


Figure 2.2: Unemployed for more than 270 days before the programme

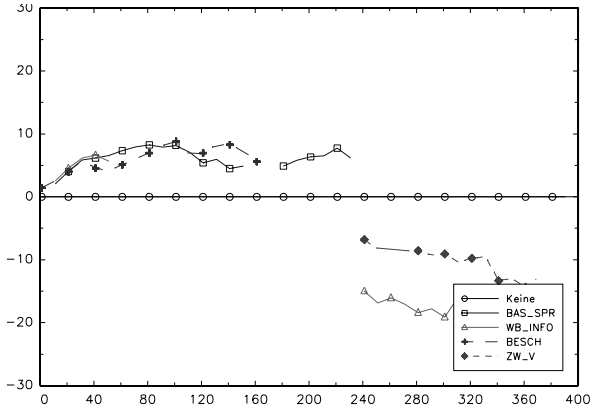


Figure 2.3: Native language not a Swiss language

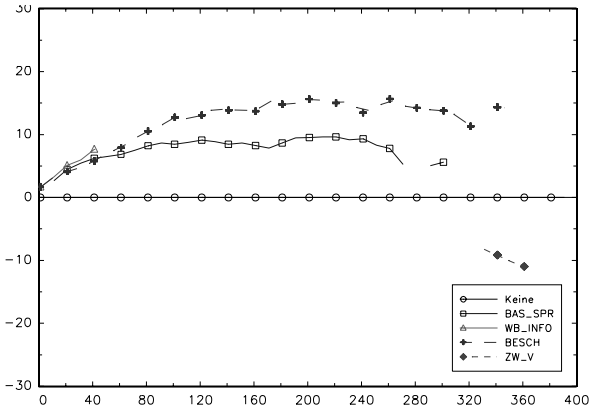
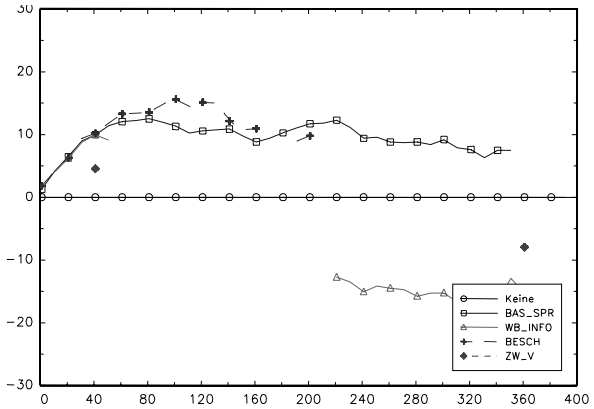


Figure 2.4: Women



Note: Day 1 corresponds to the first day after the start of the programme. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

Figure 2 presents the same effects as used for Figure 1, but changes the time perspective somewhat. The effects are now measured relative to the actual or artificial (for NONPARTICIPATION) starting dates. Generally, the findings have a very similar structure compared

to those presented in Table 1. The major difference is that there positive and quantitatively large effects for FURTHER TRAINING appear now not only for women but also for people unemployed for more than 9 months before the programme.

The major difference of the two concepts of time is that effects could differ when some programmes start systematically later, and if this fact has some influence on the treatment effect. The major issue here is probably the business cycle and perhaps the occurence of seasonal effects. An alternative concept of timing could relate the effects to the end of the programme. However, for most programmes the end date is already part of the effect, because people could leave early if they find a job. Hence, it is not an attrative concept and is not persued any further in this paper.

Using the concept of calendar time as in Figure 1, Figure 3 shows the effects of BASIC TRAINING compared to the other possible states. A positive number indicates that BASIC TRAINING increases the employment probability of the respective participants. Again Figure 3.1 covers all participants. Figure 3.2 to 3.4 relate to the specific subgroups. Taken together the results are rather negative and do not show much difference with respect to subgroups. Again, women participating in BASIC TRAINING would have obtained an optimal result if they would have participated in FURTHER TRAINING instead.

Figure 3: Effects of BASIC TRAINING for the respective participants ($\theta_0^{m,l}$ for employment)

Figure 3.1: All

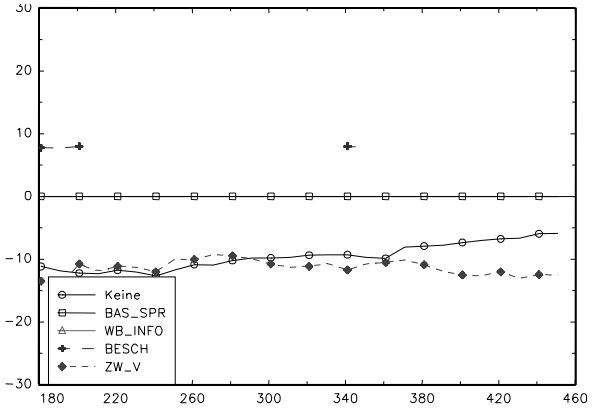


Figure 3.2: Unemployed for more than 270 days before the programme

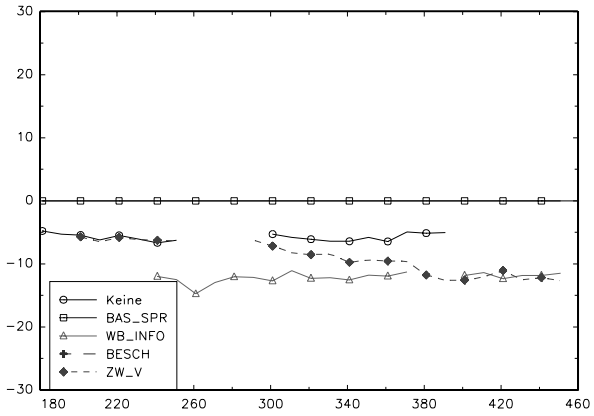


Figure 3.3: Native language not a Swiss language

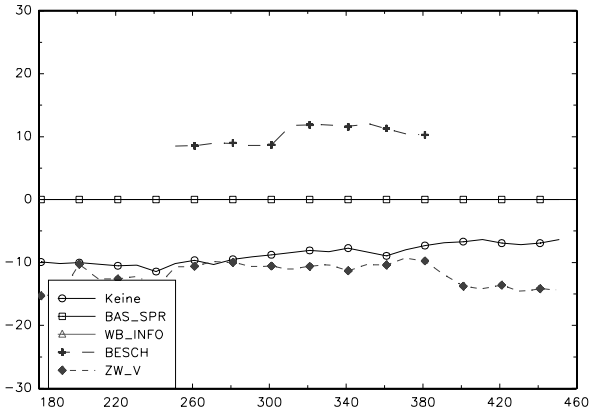
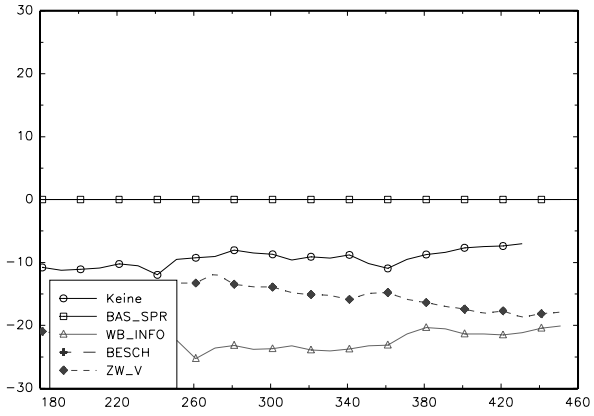


Figure 3.4: Women



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

Figure 4 presents the same type of results for FURTHER TRAINING and its participants. Basically the results confirm the view of the previous tables, although some lack of precision limits the possibilities for statistically significant comparisons.

Figure 4: Effects of FURTHER TRAINING for the respective participants ($\theta_0^{m,l}$ for employment)

Figure 4.1: All

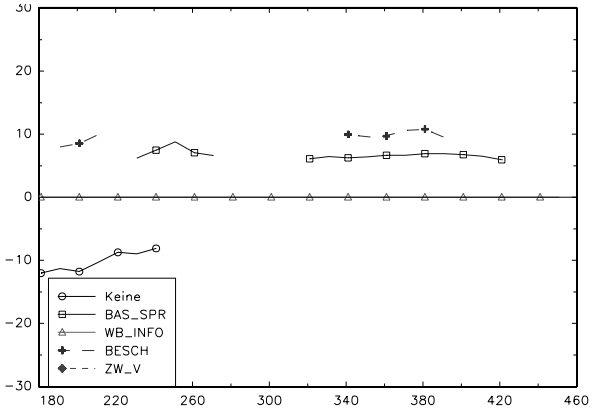


Figure 4.2: Unemployed for more than 270 days before the programme

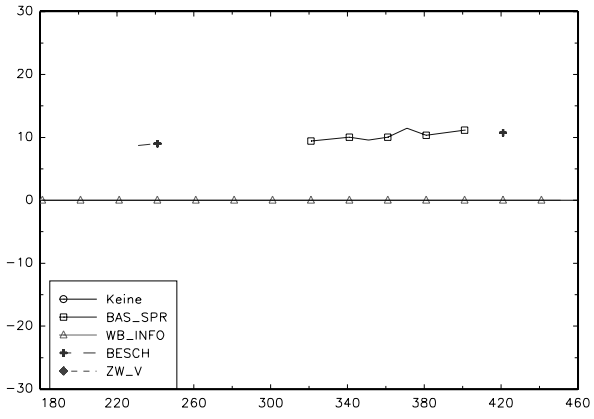


Figure 4.3: Native language not a Swiss language

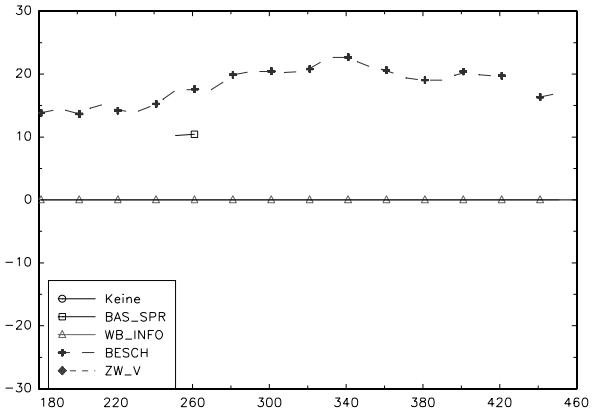
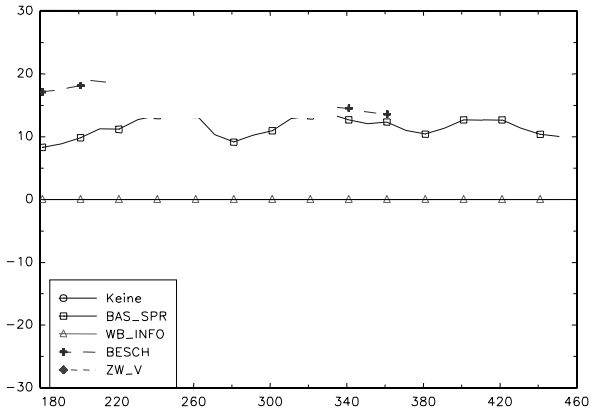


Figure 4.4: Women



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

Figure 5 confirms the view that EMPLOYMENT PROGRAMMES are dominated by many other programmes. It is interesting to note however, that for women EMPLOYMENT PROGRAMMES dominate BASIC TRAINING at least for 1999.

Figure 5: Effects of EMPLOYMENT PROGRAMMES for the respective participants ($\theta_0^{m,l}$ for employment)

Figure 5.1: All

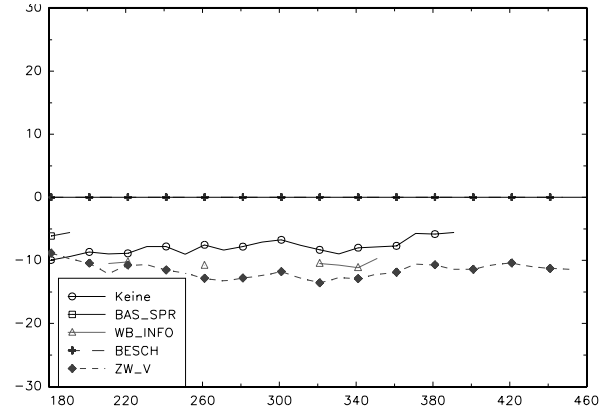


Figure 5.2: Unemployed for more than 270 days before the programme

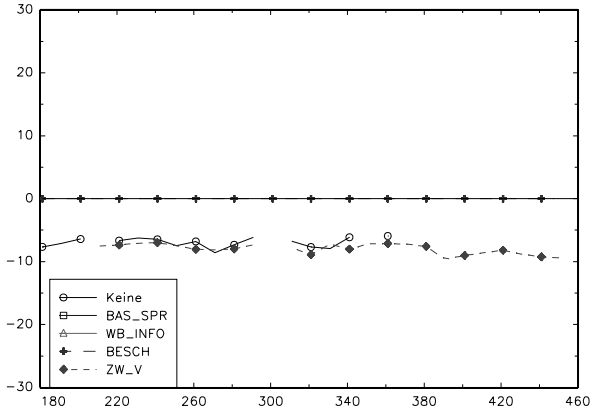


Figure 5.3: Native language not a Swiss language

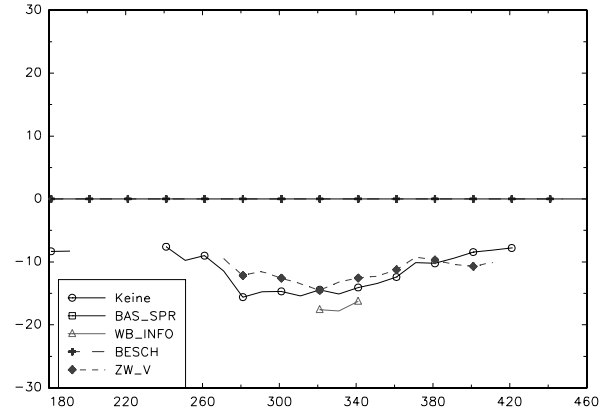
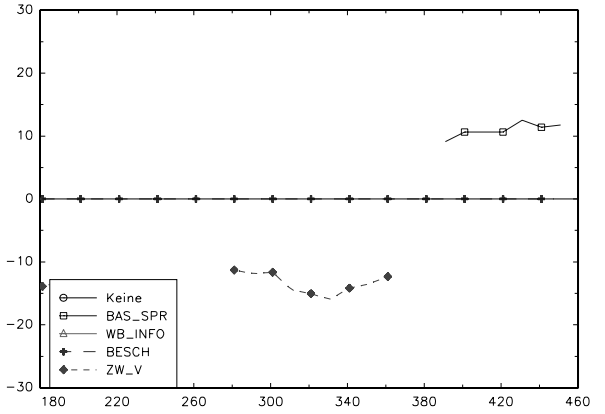


Figure 5.4: Women



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

Figure 6 shows again that TEMPORARY EMPLOYMENT is one of the more successful programmes. This holds true also for the subgroups. Again for women there are hints that FURTHER TRAINING is clearly not dominated by TEMPORARY EMPLOYMENT.

Figure 6: Effects of TEMPORARY EMPLOYMENT for the respective participants ($\theta_0^{m,l}$ for employment)

Figure 6.1: All

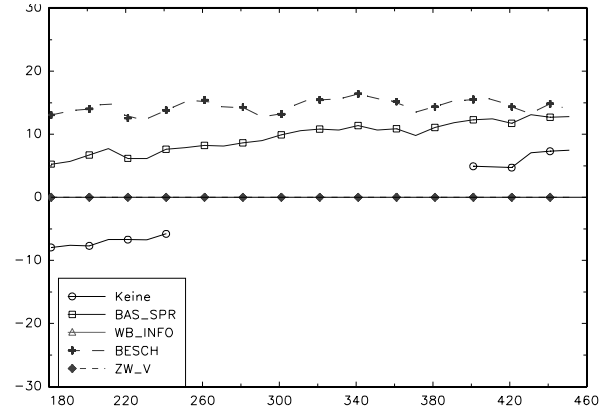


Figure 6.2: Unemployed for more than 270 days before the programme

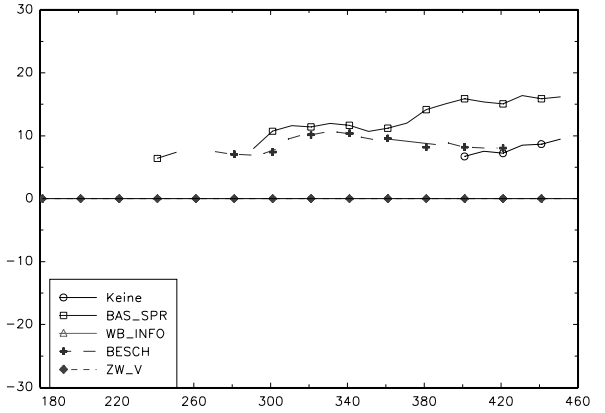


Figure 6.3: Native language not a Swiss language

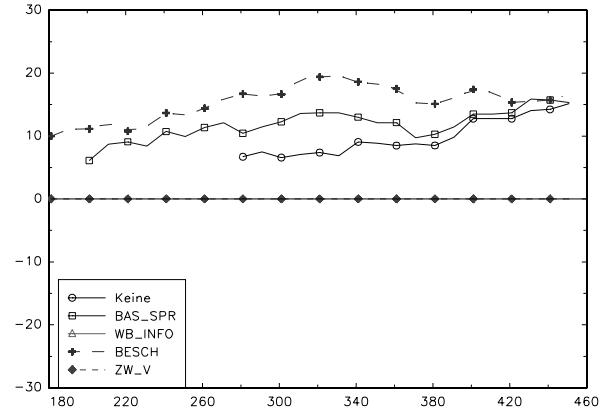
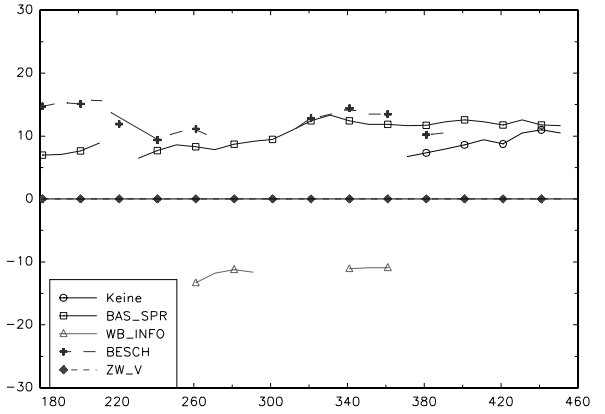


Figure 6.4: Women



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

The above considerations have shown that the proposed approach can be used to address the heterogeneity issue in many different fruitful ways. Thereby it can become a very useful tool in policy analysis. It should be emphasized again at this point that the current application serves merely as an example of what could be done. It should not be used to devise policy in the canton of Zurich.

4.5 Aggregation

Table 11 shows the aggregate effects defined in Section 2. First, considering the ones using $P(S = l \mid l \neq m)$ as weights basically confirms the ranking of the treatments that emerged from the many pair-wise comparisons. The results also confirm the a priori view that the composite effects and the effects using a binary model could be very different indeed.

Table 11: Estimation results for the composite effects

	$\hat{\gamma}_N^m(\tilde{v}^m)$	$\hat{\alpha}_N^m(\tilde{v}^m)$	$\hat{\theta}_N^m(\tilde{v}^m)$	$\hat{\theta}_N^m(\breve{v}^m)$ (aggregated)	unadjusted differences
Nonparticipation	-4.3	-4.5	-2.1	1.1	0.9
basic training	-9.5	-9.8	-6.6	-15.9	-10.8
further training	1.3	1.5	2.0	14.6	11.9
employment programme	-6.2	-6.3	-3.6	-26.7	-13.7
temporary employment	10.6	10.7	9.2	27.0	12.7

Note: The outcome variable is *employed* for day 451 (in %-points). The first three columns are computed from *the MVP unconditional* estimates. The effects presented in the last but one column are computed by aggregating the respective non-treatment groups before the estimation of the effect.

Figure 7 displays some dynamic aspect of the most interesting two aggregate measure, namely $\hat{\gamma}_N^m(\tilde{v}^m)$ and $\hat{\theta}_N^m(\tilde{v}^m)$. The differences between these two measures appear to be small. The continuous downward drift of NONPARTICIPATION in the relative ranking is again a striking feature of these figures. This suggests a possibly larger difference between the short term and the long term effects of programme participation.

Figure 7: Aggregate effects of the various programmes over time (employment)

Figure 7.1: $\hat{\gamma}_N^m(\tilde{v}^m)$

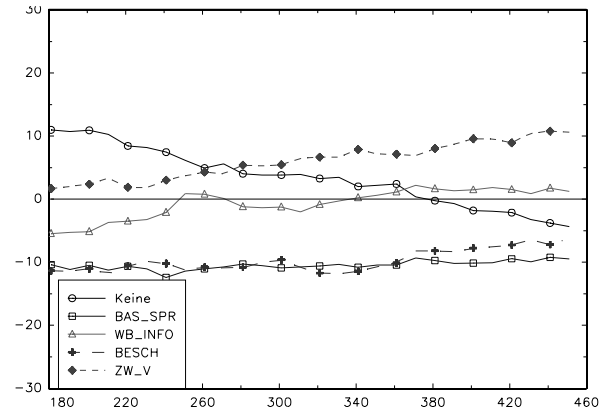
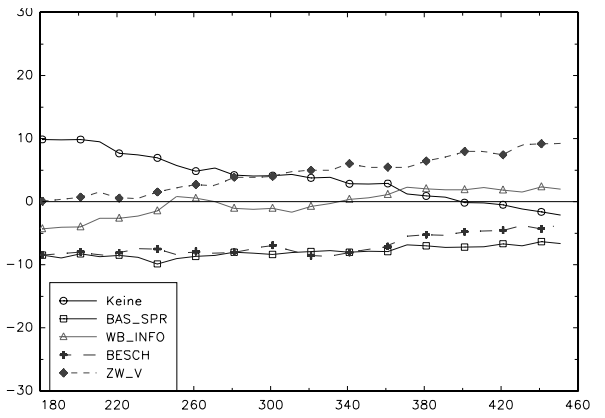


Figure 7.2: $\hat{\theta}_N^m(\tilde{v}^m)$



Note: Day 366 corresponds to the beginning of 1999. *Keine*: Nonparticipation; *BAS_SPR*: Basic Training; *WB_INFO*: Further training; *BESCH*: Employment programme; ZW_V: Temporary employment. A positive number means that participation in the respective programme increased the employment probability compared to being in one of the other states. The figures display only values that are significant at the 5% level. See also note below Table 10.

The effect for the subgroups suggest on the one hand very large effects of TEMPORARY EMPLOYMENT, in particular for individuals with a foreign native language. Again the ranking of the programmes appear to be different for women than for men. In particular TEMPORARY EMPLOYMENT does appear as the single positive programme, but FURTHER TRAINING and even EMPLOYMENT PROGRAMMES seem to come close to it as well, at least in 1999.

Figure 8: Aggregate effects for subgroups

Figure 8.1: $\hat{\gamma}_N^m(\tilde{v}^m)$: Unemployed for at least 270 days before the programme

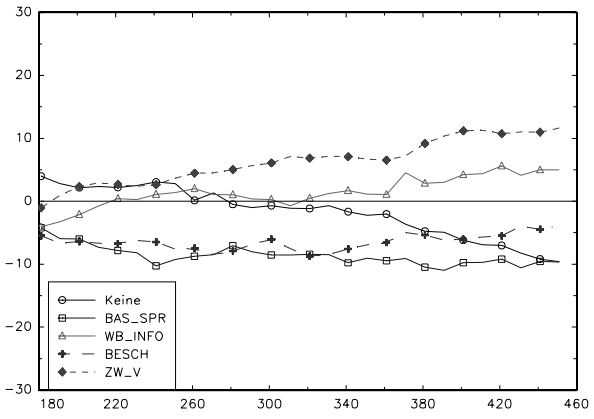


Figure 8.2: $\hat{\theta}_N^m(\tilde{v}^m)$: Unemployed for at least 270 days before the programme

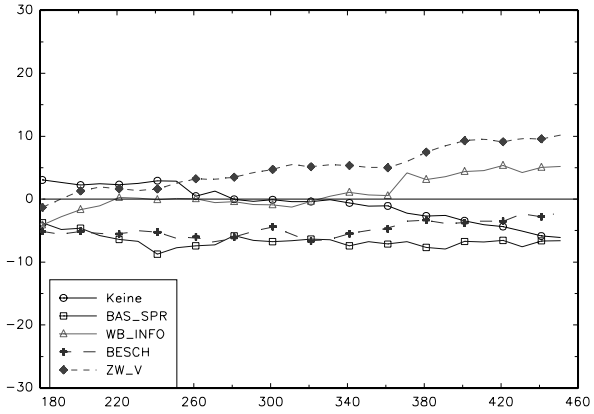


Figure 8.3: $\hat{\gamma}_N^m(\tilde{v}^m)$: Native language not Swiss

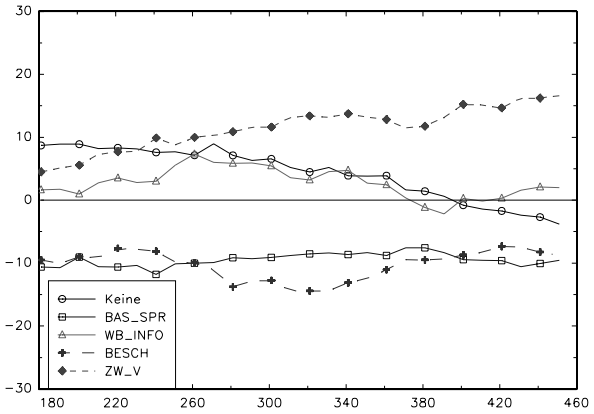


Figure 8.4: $\hat{\theta}_N^m(\tilde{v}^m)$: Native language not Swiss

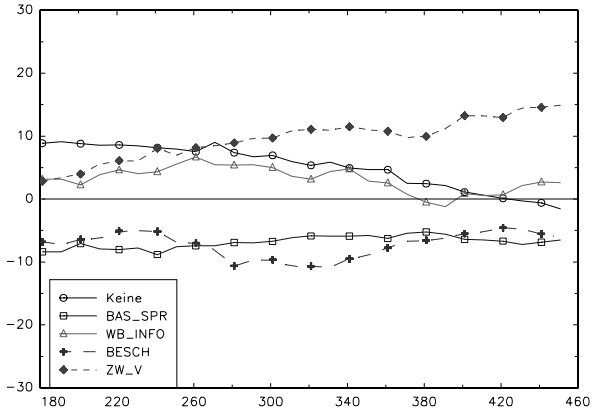


Figure 8.3: $\hat{\gamma}_N^m(\tilde{v}^m)$: Women

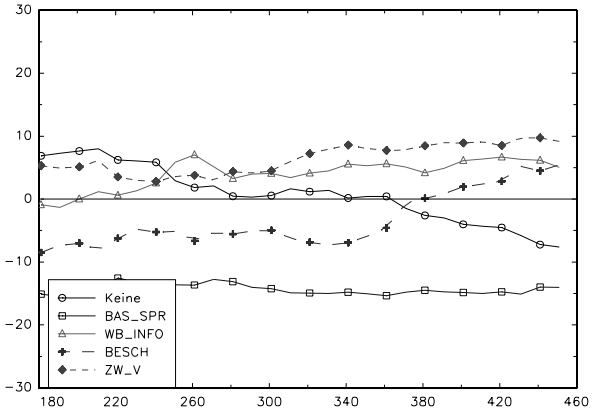
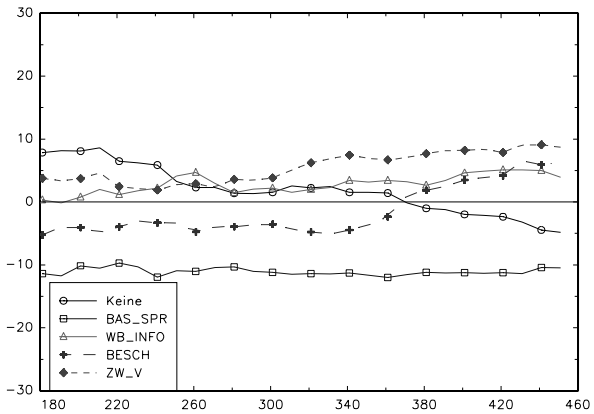


Figure 8.5: $\hat{\theta}_N^m(\tilde{v}^m)$: Women



See note under Figure 7.

5 Conclusion

The paper suggests an approach of handling the issue of multiple treatments in microeconometric evaluation studies based on balancing score matching. The proposed methods have been applied as an example to the evaluation of active labour market policies. The application compared different estimators in practise and showed that the multiple treatment approach can lead to valuable insights that might be lost otherwise. It gave also some hints on the interpretation of the different effects computed by aggregating different treatments.

The paper also showed that there are two different approaches to modelling the respective balancing scores needed for matching. The first approach is to derive the probabilities used for the balancing scores by specifying and estimating a multiple discrete choice model. The alternative is to concentrate on modelling and estimating (all) directly conditional probabilities between any two choice. One advantage of using a multinomial discrete choice model instead of concentrating only on the binary conditional choices is that using this approach it is easier to understand the empirical contents of the joint selection process. The drawback is that it is computational much more expensive. Furthermore, there is a lack of robustness, in the sense that a misspecification of one choice equation could lead to inconsistent estimates of all conditional choice probabilities. In the application the simultaneous approach appeared to be superior. However, considerable future research is needed to see whether this result holds in other circumstances as well and how this results depends on the particular specification of the multinomial choice model and the measurement of the differences between the two multidimensional distributions.

Appendix A: Technical Appendix

The first part of this appendix contains the proofs that the composite effects $\gamma_0^m(v^m)$ and $\theta_0^m(v^m)$ have a causal interpretation in terms of the composite potential outcome $Y^{-m} = \sum_{l=0}^{M} v^{m,l} Y^l$.

$$\gamma_0^m(v^m) = \sum_{l=0}^{M} v^{m,l} \gamma_0^{m,l} = \sum_{l=0}^{M} v^{m,l} [E(Y^m) - E(Y^l)]$$

$$= E(Y^m - \sum_{l=0}^{M} v^{m,l} E(Y^l)$$

$$= E(Y^m) - E(\sum_{l=0}^{M} v^{m,l} Y^l)$$

$$= E(Y^m) - E[Y^{-m}(v^m)]. \qquad \text{q.e.d.}$$

The same line of argument is valid for $\theta_0^m(v^m)$ as well:

$$\theta_0^m(v^m) = \sum_{l=0}^{M} v^{m,l} \theta_0^{m,l} = \sum_{l=0}^{M} v^{m,l} [E(Y^m \mid S = m) - E(Y^l \mid S = m)]$$

$$= E(Y^m \mid S = m) - \sum_{l=0}^{M} v^{m,l} E(Y^l \mid S = m)$$

$$= E(Y^m \mid S = m) - E[(\sum_{l=0}^{M} v^{m,l} Y^l) \mid S = m]$$

$$= E(Y^m \mid S = m) - E[Y^{-m}(v^m) \mid S = m]. \qquad \text{q.e.d.}$$

Furthermore, note such an interpretation appears not to be available for $\alpha_0^m(v^m)$:

$$\alpha_0^m(v^m) = \sum_{l=0}^{M} v^{m,l} \alpha_0^{m,l} = \sum_{l=0}^{M} v^{m,l} [E(Y^m \mid S = m, S = l) - E(Y^l \mid S = m, S = l)].$$

Since the conditioning sets depend on the index of the summation operator, a further simplification is not possible.

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 A. B. Krueger (1999): "Empirical Strategies in Labor Economics", forthcoming in O. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. III.

Börsch-Supan, A. and Hajivassiliou, V.A. (1993): "Smooth Unbiased Multivariate Probabilities Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models", *Journal of Econometrics*, 58, 347-368.

Gerfin, M., and M. Lechner (1999): "Evaluating the Swiss Active Labour Market Policies Using Microeconometrics: The Report to the Government", unpublished, in German, paper not yet available.

Geweke, J., M. Keane, D. Runkle (1994): "Alternative Computational Approaches to Inference in the Multinomial Probit Model", *Review of Economics and Statistics*, 1994, 609-632.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", forthcoming in O. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. III.

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

Imbens, G. W. (1999): "The Role of the Propensity Score in Estimating Dose-Response Functions", *NBER Technical Working Paper*, 237, 1999.

Lechner, M. (1999a): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification", *Journal of Business & Economic Statistics*, 17, 74-90.

Lechner, M. (1999b): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption", *Discussion paper 9908, University of St. Gallen*.

McFadden, D. (1984): "Econometric Analysis of Qualitative Response Models", in Z. Griliches and M. D. Intriligator (editors), *Handbook of Econometrics*, Volume 2, 1396-1457.

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.

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.

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 to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Rubin, D. B. (1991): "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism", *Biometrics*, 47, 1213-1234.