PROGRAM HETEROGENEITY AND PROPENSITY SCORE MATCHING: AN APPLICATION TO THE EVALUATION OF ACTIVE LABOR MARKET POLICIES

Michael Lechner*

Abstract—This paper addresses microeconometric evaluation by matching methods when the programs under consideration are heterogeneous. Assuming that selection into the different subprograms and the potential outcomes are independent given observable characteristics, estimators based on different propensity scores are compared and applied to the analysis of active labor market policies in the Swiss region of Zurich. Furthermore, the issues of heterogeneous effects and aggregation are addressed. The results suggest that an approach that incorporates the possibility of having multiple programs can be an informative tool in applied work.

I. Introduction

There is a considerable discrepancy between technically sophisticated modern microeconometric evaluation methods and real programs to be evaluated when it comes to taking account of program heterogeneity. Standard microeconometric evaluation methods are mostly concerned with the effects of being or not being in a particular program, whereas, for example in active labor market policies (ALMP), there is usually a range of heterogeneous subprograms, such as training, public employment programs, or job counseling.[1] These subprograms often differ with respect to their target population, their contents and duration, their selection rules, and their effects.

When participation in such programs is independent of the subsequent outcomes conditionally on observable exogenous factors (conditional independence assumption (CIA)), the standard model of only two states—that is, participation versus nonparticipation—is extended by Imbens (1999) and Lechner (2001a) to the case of multiple states ("treatments").[2] Both papers show that the important dimension-reducing device of the binary treatment model, called the *balancing score property of the propensity score,* is still valid in principle, but needs to be suitably revised.

Here, several estimation methods suitable in that framework, all based on matching on the propensity score, are compared and applied to the evaluation of active labor market policies in the Swiss canton of Zurich. The aim of this study, which is one of the first empirical implementations of this approach, is to give an example of how an evaluation could be performed in this setting.[3] The comparison of the performance of the different estimators in practice provides information relevant for other studies. In addition, the application shows that the multiple-treatment approach can lead to valuable insights. It is, however, beyond the scope of this paper to derive policy-relevant conclusions.

The paper is organized as follows. 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 independence assumption. Section III proposes matching estimators for this setting. Section IV presents the empirical baseline results for the Swiss region of Zurich. Section V investigates more on the issue of effect heterogeneity and section VI more on aggregation. In the latter, a causal parameter is developed that corresponds to a comparison of a specific treatment to a composite state that is composed of an aggregation of the remaining states. Section VII concludes. Appendix A discusses technical details concerning aggregation, and appendix B presents the results of a multinomial probit estimation for the participation in the different states.

II. The Causal Evaluation Model with Multiple Treatments

A. Notation and Definition of Causal Effects

In the prototypical model of the microeconometric evaluation literature, an individual faces two states of the world, such as participation in a training program or nonparticipation in such a program. She gets a hypothetical (potential) outcome for both states, and the causal effect is defined as

Received for publication December 3, 1999. Revision accepted for publication March 20, 2001.

*Swiss Institute for International Economics and Applied Economic Research.

I am also affiliated with CEPR, London; ZEW, Mannheim; and IZA, Bonn. Financial support from the Swiss National Science Foundation (projects 12-53735.18, 4043-058311, and 4045-050673) is gratefully acknowledged. The data are a subsample from a database generated for the evaluation of the Swiss active labor 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. This paper has been presented at the Evaluation of Labor Market Policies workshop, Bundesanstalt für Arbeit (IAB), in Nuremberg, 1999, as well as at the annual meeting of the population economics section of the German Economic Association in Zurich, 2000. I thank participants for helpful comments and suggestions. Furthermore, I thank two anonymous referees of this journal for critical but very helpful remarks on a previous version. I also thank Heidi Steiger for carefully reading the manuscript. All remaining errors are my own.

[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 programs 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, because, for example, an employment program offered in two different possible lengths (the doses) could always be redefined as being two separate programs.

[3] Brodaty, Crepon, and Fougère (2001) and Larsson (2000) are further applications based on this approach.

difference of these potential outcomes. This model is known as the Roy (1951)–Rubin (1974) model (RRM).[4]

Consider now a world with $(M + 1)$ mutually exclusive states. (The states are also called *treatments* in the following text to preserve the terminology of that literature.) The potential outcomes are denoted by $\{Y^0, Y^1, \ldots, Y^M\}$. For every person, a realization from only one element of $\{Y^0, Y^1, \ldots, Y^M\}$ is observable. The remaining M outcomes are *counterfactuals* in the language of RRM. Participation in a particular treatment is indicated by the variable $S \in \{0, 1, \ldots M\}$.

To account for the $(M + 1)$ possible treatments, the definitions of average treatment effects developed for binary treatments need to be adjusted.[5] Here, the focus is on a pairwise comparison of the effects of treatments m and l for the participants in treatment m. This is the multiple-treatment version of the average treatment effect on the treated, which is the parameter typically estimated in evaluation studies:[6]

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

$\theta_0^{m,l}$ denotes the expected effect for an individual randomly drawn from the population of participants in treatment m ($\theta_0^{m,m} = 0$).[7] Note that, if the effects of participants in treatments m and l differ for the two subpopulations participating in m and l, respectively, then the treatment effects on the treated are not symmetric ($\theta_0^{m,l} \neq -\theta_0^{l,m}$).

B. Identification

RRM clarifies that the average causal treatment effect is generally not identified. Identification is obtained by untestable assumptions. Their plausibility depends on the substance of the economic 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)).

Imbens (1999) and Lechner (2001a) consider identification under the multiple-treatment version of CIA that states that all potential treatment outcomes are independent of the assignment mechanism for any given value of a vector of attributes, X, in an attribute space, χ. They show that CIA identifies the parameters of interest. CIA is formalized in expression (2), in which $\perp\!\!\!\perp$ denotes independence:

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

Assume also the common support condition to be valid, that is, that for all $x \in \chi$, there is a positive probability of every treatment to occur.[8] CIA requires the researcher to observe all characteristics that jointly influence the potential outcomes as well as the selection into the treatments.[9] In that sense, CIA may be called a "data hungry" identification strategy.

Rubin (1977) and Rosenbaum and Rubin (1983) show for the binary treatment framework that it is in fact not necessary to condition on the attributes, but only to condition on the participation probability conditional on these attributes (propensity score). Thus, the dimension of the estimation is reduced, given a consistent estimate of the propensity score.

Imbens (1999) and Lechner (2001a) show that properties similar to the propensity score property hold in a multiple-treatment framework as well. For the average treatment effect on the treated specifically, Lechner (2001a, proposition 3) shows the following:

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

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

$\theta_0^{m,l}$ is identified from an infinitely large random sample, because all participation probabilities, as well as $E(Y^m | S = m)$ and $E(Y^l | P^{l|ml}(X), S = l)$, are identified. The dimension of the estimation problem is reduced to one. This result suggests that usual nonparametric methods (those used in the binary treatment framework) that condition on an estimated propensity score can be applied here as well.

A corollary of this result is that, to identify $\theta_0^{m,l}$, only information from the subsample of participants in m and l is needed. However, for example, when all values of m and l are of interest, then all the sample is needed for identification. Even in this case, one may still model and estimate the $M(M - 1)/2$ binary conditional probabilities $P^{l|ml}(x)$.

It may be more straightforward from a modeling point of view to model the individual simultaneous discrete-choice problem involving all states. $P^{l|ml}(x)$ could then be computed from that model.[10] When such a discrete-choice

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

[5] Assume for the rest of the paper that the typical assumptions of the RRM are fulfilled. (See Holland (1986) or Rubin (1974) for example.) Particularly, these assumptions rule out dependence or interference between individuals.

[6] In section IV, other effects that correspond in some sense to the average treatment effects for the population in the binary case are considered as well.

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

[8] This version of the common support condition is in fact unnecessarily restrictive. The precise version is given by Lechner (2001a). Furthermore, Lechner (2001b) discusses violations of the common support condition and establishes informative bounds for the effects when such violations occur. These issues are beyond the scope of this paper.

[9] Note that CIA can be seen as too restrictive because only conditional mean independence (CMIA) is needed to identify mean effects. However, CIA has the virtue that, with CIA, CMIA is valid for all transformations of the outcome variables. Furthermore, in many applications, it is usually difficult to argue why CMIA holds and CIA is violated.

[10] $P^{l|ml}(x) = P^l(x)/[P^l(x) + P^m(x)]$; $P^l(x) := P(S = l | X = x)$.

TABLE 1.—A MATCHING PROTOCOL FOR THE ESTIMATION OF $\theta_0^{m,l}$

Step 1	Estimate the propensity score. a) Either specify and estimate a multinomial choice model to obtain $[\hat{P}_N^0(x), \hat{P}_N^1(x), \ldots, \hat{P}_N^M(x)]$; compute $\hat{P}_N^{l\vert ml}(x) = \frac{\hat{P}_N^l(x)}{\hat{P}_N^l(x) + \hat{P}_N^m(x)}$. b) or specify and estimate the conditional probabilities on the subsample of participants in m and l for all different combinations of m and l to obtain $\tilde{P}_N^{l\vert ml}(x)$.
Step 2	Estimate the expectations of the outcome variables conditional on the respective propensity scores for m and l. For a given value of m and l, the following steps are performed: a) Choose one observation in the subsample defined by participation in m and delete it from that subsample. b) Find an observation in the subsample of participants in l that is as close as possible to the one chosen in step 2(a) in terms of $\hat{P}_N^{l\vert ml}(x)$, $\tilde{P}_N^{l\vert ml}(x)$ or $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$. If using the multivariate score $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$, "closeness" is based on the Mahalanobis distance. The weighting matrix is the inverse covariance matrix of $[\hat{P}_N^m(x), \hat{P}_N^l(x)]$ in the pool of participants in l. Do not remove that observation, so it can be used again. c) Repeat (a) and (b) until no participant is left in subsample m. d) Using the matched comparison group formed in (c), compute the respective conditional expectation $(E(Y^l\vert S = m))$ by the weighted sample mean $\hat{E}_N(Y^l\vert S = m)$. Note that the same observations may appear more than once in that group and thus have different weights corresponding to the number of their occurrence in the respective comparison sample. Compute the estimate of $E(Y^m\vert S = m)$ as sample mean in subsample of participants in m $\hat{E}_N(Y^m\vert S = m)$. e) Compute the variance of $\hat{E}_N(Y^l\vert S = m)$ by $\sum_{i\in l} (\hat{w}_i^{m,l})^2/(N^m)^2 \widehat{\mathrm{Var}}_N(Y\vert S = l)$ and the variance of $\hat{E}_N(Y^m\vert S = m)$ by $\widehat{\mathrm{Var}}_N(Y\vert S = m)/N^m$. $\widehat{\mathrm{Var}}_N(Y\vert S = j)$ denotes the empirical variance in the respective subpopulation, N^m denotes the number of participants in m, and $\hat{w}_i^{m,l}$ denotes the number of times observation i who is a participant in l appears in the control group formed to estimate $\hat{E}_N(Y^l\vert S = m)$.
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 as $\hat{\theta}_N^{ml} = \hat{E}_N(Y^m\vert S = m) - \hat{E}_N(Y^l\vert S = m)$. The corresponding variances are given by the sum of $\widehat{\mathrm{Var}}_N(Y\vert S = m)/N^m$ and $\sum_{i\in l} (\hat{w}_i^{m,l})^2/(N^m)^2 \widehat{\mathrm{Var}}_N(Y\vert S = l)$.

The estimator of the asymptotic standard error of $\hat{\theta}_N^{ml}$ is based on the approximation that the estimation of the weights can be ignored. Using bootstrap to obtain an estimate of the distribution of $\hat{\theta}_N^{ml}$ is an alternative explored by Lechner (2000b). It turned out that the approximate standard errors are somewhat too small, but not by much. Due to the computational expense of the multinomial probit with five categories and four hundred draws in the GHK simulator as used in the following application, bootstrap quantiles of the estimated effects are not provided.

model is estimated or generally when the conditional choice probabilities are more difficult to obtain than the marginal ones, it could be attractive to condition jointly on the two marginal probabilities, $P^l(X)$ and $P^m(X)$, instead of $P^{l\vert ml}(X)$. Conditioning on $P^l(X)$ and $P^m(X)$ also identifies $\theta_0^{m,l}$ because $P^l(X)$ together with $P^m(X)$ is finer than $P^{l\vert ml}(X)$ (meaning that $P^{l\vert ml}(X)$ is the same as its expectation conditional on $P^l(X)$ and $P^m(X)$):

$$E[P^{l\vert ml}(X)\vert P^l(X), P^m(X)] = E\left[\frac{P^l(X)}{P^l(X) + P^m(X)}\middle\vert P^l(X), P^m(X)\right] = P^{l\vert ml}(X). \quad (4)$$

III. A Matching Estimator

Given the choice probabilities or a consistent estimate of them, the terms appearing in equation (3) can be estimated by any parametric, semiparametric, or nonparametric regression method that can handle one- or two-dimensional explanatory variables. In many cases, CIA is exploited using a matching estimator; for recent examples, see Angrist (1998), Dehejia and Wahba (1999), Heckman, Ichimura, and Todd (1998), and Lechner (1999), among others.

For the multiple-treatment model, Lechner (2001a) proposes a matching estimator that is as analogous as possible to the rather simple algorithms used in the literature on binary treatment evaluation. (See table 1.)

Note that this implementation of matching allows the same comparison observation to be used repeatedly. This modification is necessary for the estimator to be at all applicable when the number of participants in treatment m is larger than in the comparison treatment l because the role of m and l as treatment and control is reversed during the estimation. This procedure has the potential problem that very few observations may be heavily used, although other very similar observations are available, leading to an unnecessary inflation of variance. Therefore, the occurrence of this feature should be checked, and, if it appears, the algorithm needs to be suitably revised.[11] Similar checks need to be performed—as usual—to make sure that the distributions of the balancing scores overlap sufficiently in the respective subsamples. For subsamples m and l, this condition means that the distributions of $\hat{P}_N^{l\vert ml}(x)$ (or $\tilde{P}_N^{l\vert ml}(x)$ 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. However, it is not asymptotically efficient because the typical tradeoff appearing in nonparametric regression between bias and variance is not addressed. (It is actually minimizing the bias.) Other more sophisticated and more computer-intensive matching methods are discussed for example by Heckman, Ichimura, and Todd (1998).[12]

[11] In that case, a simple alternative would be to use the "blocking" approach suggested by Rosenbaum and Rubin (1985).

[12] Note that algorithms like kernel smoothing could be asymptotically more efficient. However, to compare binary and multiple treatments, it appears advisable to use commonly used and stable algorithms and to avoid discussions about optimal bandwidth choice and other issues akin to the asymptotically more efficient methods. For a comparison of the various nonparametric methods, see Frölich (2000).

TABLE 2.—DESCRIPTIVE STATISTICS OF SELECTED VARIABLES FOR SUBSAMPLE DEFINED BY DIFFERENT STATES

	No Participation	Basic Training	Further Training	Employment Program	Temporary Wage Subsidy
	Median in Subsample				
Age	39	38	40	40	39
Days of unemployment before start of program	251	218	219	335	247
Duration of program in days	0	63	41	155	113
Starting day (1 corresponds to 1/1/97)	89	82	76	156	107
	Share in Subsample in %				
Gender: female	46	56	43	37	43
Subjective valuations of labor 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

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

IV. Empirical Application

A. *Introduction and Descriptive Statistics*

After experiencing increasing rates of unemployment in the mid-1990s, Switzerland conducted a substantial active labor market policy with several different subprograms. For the purpose of this study, they are aggregated into five different groups of more or less similar states: NO PARTICIPATION in any program, BASIC TRAINING (including job counseling and courses in the local language), FURTHER vocational TRAINING (including information technology courses as the most important part), EMPLOYMENT PROGRAMS, and a TEMPORARY WAGE SUBSIDY (job with company at a lower wage, with the labor office paying the difference between the wage and 70%–80% of previous earnings[13]).

This application concentrates only on the largest Swiss canton, Zurich.[14] The data originate from the Swiss unemployment registers and cover the population unemployed in the canton of Zurich. After selection, it covers persons unemployed on December 31, 1997 (unemployment is a condition for eligibility), aged between 25 and 55, who have not participated in a program before the end of 1997 and are not disabled. Individual program participation begins during 1998 and the observation period ends in March 1999. Further information about the database can be found in Gerfin and Lechner (2000).[15] The database is fairly informative because it contains all the information that the local labor 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.[16]

Table 2 shows descriptive statistics of selected variables for subsamples defined by the five different states. From these statistics, it is obvious that there is heterogeneity with respect to program characteristics, such as duration, as well as with respect to characteristics of participants such as skills, qualifications, employment histories, among others.[17]

[13] The unemployed receives slightly more money than unemployment benefits. Furthermore, the expiration date of unemployment benefits may be prolonged.

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

[15] Gerfin and Lechner (2000) study the effects of the various programs of the Swiss active labor market policy. Their database 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.

[16] Obviously, there may be substantial arguments claiming that this may not be true. However, the aim of this study is to provide an example of how an evaluation could be performed in this setting, not to derive policy-relevant conclusions. The reader is referred to Gerfin and Lechner (2000) for more discussion about the features of the programs as well as the selection rules. They address also the issue whether there might be additional unobserved factors correlated with outcomes and selection that could invalidate the CIA.

[17] Unemployment duration until the beginning of training is an important variable for the participation decision. Because 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 closely follows an approach called *random* by Lechner (1999a). Alternative approaches are discussed by Lechner (1999a, 2000b).

TABLE 3.—UNADJUSTED DIFFERENCES AND LEVELS OF EMPLOYMENT IN %-POINTS

	No Participation	Basic Training	Further Training	Employment Program	Temporary Wage Subsidy	All Other Categories
No participation	(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 program				(25.8)	−22.7	−13.7 (38.3)
Temporary wage subsidy					(48.5)	12.7 (35.8)

The outcome variable is *employment* in percentage points for day 451 (end of March 1999). Absolute levels on main diagonal and in the last column (in brackets). All Other Categories denotes the aggregation of all categories except the one given in the respective row.

The effects of the programs are measured in terms of changes in the average probabilities of employment in the first labor market caused by the program after the program begins. The time in the program is not considered as regular employment. 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 use the latest observations available, those of 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.

The results show a wide range of average employment rates. The highest values that are close to 50% correspond to FURTHER TRAINING and TEMPORARY WAGE SUBSIDY. Clearly, the participants with the worst postprogram employment experience are participants in EMPLOYMENT PROGRAMS, followed by participants in BASIC TRAINING. However, it is yet impossible to decide whether the resulting order of employment rates is due to different effects of the programs or to a systematic selection of unemployed with fairly different employment chances into specific programs. Disentangling these two factors is the main task of every evaluation study.

B. *Participation Probabilities*

Section III showed that the participation probabilities are major ingredients for the matching estimator. Beyond that (direct) purpose, an empirical analysis of the participation decision may also reveal information about the selection process that could not be obtained from an analysis of the institutions alone, and that may be an important piece of information on its own—particularly so if it turns out that the effect of the programs are heterogeneous and that this heterogeneity is correlated with variables appearing prominently in the selection process. This issue is considered in more detail in section V.

From the point of view of using the selection probabilities as input to the matching estimator, there are the two already mentioned possibilities: modeling and estimating each conditional binary choice equation separately to obtain $P^{l|ml}(x)$, for example by a binary probit or logit model, could be called a reduced-form approach. This estimation is confined to observations being in either state m or l. Thus, it closely mirrors the typical propensity score approach for binary treatments. The only difference is that it has to be performed $M(M-1)/2$ times on different subsamples to obtain all necessary probabilities. It does not impose the "independence of irrelevant alternative" assumption. For the current application, ten equations are estimated. Obviously, issues such as documentation of the results, monitoring variable selection and quality of the specification, checking the common support condition, and the interpretation of the results becomes very tedious. Although ten binary probits are still possible in the current application with only five categories, for papers that perform a more disaggregated analysis the reduced-form approach becomes prohibitive.[18]

The alternative to the reduced-form approach could be called *structural approach*. The idea is to formulate the complete choice problem in one model and estimate it on the full sample. Popular models for such an exercise are multinomial logit (MNL) or probit (MNP) models. Both models, as well as others, can be motivated by the random utility maximization approach (McFadden, 1981, 1984). Compared to the MNL, the MNP has the advantage that it is more flexible, because it does not require the independence of irrelevant alternatives assumption to hold.[19] The estimated marginal probabilities or conditional probabilities derived from that model can then be used as input to matching. Note that the terms *reduced-form approach* and *structural approach* are imprecise, because, for example, when binary and multinomial probits are used, both approaches are not parametrically nested and the covariates influence the conditional probabilities in different functional forms. Thus, it is not possible to recover the structural parameter from the reduced-form parameters. Nevertheless, it is fair to say that the MNP appears to be (approximately) more restrictive because it is based on fewer coefficients and the derived conditional probabilities are interdependent. (Thus, the MNP structure imposes restrictions on the derived binary conditional probabilities that may be implied by a direct estimation of that probability.) Thus, contrary to

[18] For example, Gerfin and Lechner (2000) consider the case of $M = 9$. Clearly, taking sensible care of 36 probits would be very difficult. In addition, given current page limits, no journal would be prepared to publish the results of 36 probits anyway (and no reader would read them, even if the results were published).

[19] In practice, 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 appendix B.)

TABLE 4.—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 Program			Temporary Wage Subsidy		
Samples	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
No participation	6	21	49	1	8	23	1	6	25	8	18	31
Basic training	*11*	*32*	*69*	1	5	23	1	5	23	5	16	29
Further training	5	18	41	*4*	*14*	*27*	1	5	23	10	19	32
Employment program	5	17	45	1	6	20	*3*	*15*	*36*	10	20	34
Temporary wage subsidy	5	20	49	1	8	23	1	8	27	*11*	*21*	*37*
All	6	23	56	1	8	23	1	7	26	7	18	32
	Correlation Matrix of Probabilities in Full Sample											
No participation		−0.48			0.10			−0.24			−0.12	
Basic training					−0.47			−0.37			−0.56	
Further training								−0.22			0.12	
Employment program											0.18	

Based on estimation results presented in appendix B. NO PARTICIPATION is the reference category in the MNP estimation.

the reduced-form approach, if one choice equation is misspecified, all conditional probabilities could be misspecified. Another advantage of the reduced-form approach is that it avoids the cumbersome estimation of the MNP model and the choices necessary in specifying the MNP.[20] The comparison of the performance of both approaches is one of the topics of this paper.

Details of the estimation of the MNP using simulated maximum likelihood are given in appendix B. Because the substantive results of that estimation are not of primary interest for this paper, only a few remarks follow. The largest group (NO PARTICIPATION) is chosen as the reference category, and the variables 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 and on a preliminary estimation of the MNP, the final specification contains variables that describe attributes related to personal characteristics, valuations of individual skills and chances on the labor market as assessed by the labor office, previous and desired future occupations, as well as information related to the current and previous unemployment spell. Compared to the status NO PARTICIPATION, the estimated coefficients are fairly heterogeneous across choice equations, including sign changes of significant variables. Thus, the MNP confirms again the heterogeneity of the selection process. It also shows that heterogeneity is related to more variables than just those given in table 2. The results confirm that individuals with severe problems on the labor market have a higher probability of ending up in either BASIC TRAINING or an EMPLOYMENT PROGRAM. Participation in the latter is particularly likely for the long-term unemployed. The unemployed with better chances on the labor market are more likely to participate in either FURTHER TRAINING or TEMPORARY WAGE SUBSIDY. Therefore, various groups of active labor market policies are targeted to different groups of unemployed.

The estimation results of the MNP are used to compute the marginal participation probabilities of the various categories conditional on X. Table 4 shows descriptive statistics of the distribution of these probabilities in the various subgroups. The columns of the upper part of the 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 category in which these observations are observed (shown in italic) are the highest one in each column. The probabilities vary considerably. Hence, observations participating in the same treatment show a considerable heterogeneity with respect to their characteristics. This implies that there is probably sufficient overlap as is necessary for the successful working of matching and every other nonparametric estimator.[21]

The lower part of table 4 presents the correlations of these probabilities in the sample. There are fairly strong negative correlations between the probabilities for some treatments, but they are not less 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 (not given here).

For the reduced-form approach, ten binary probit models using the variables appearing in the corresponding two choice equations of the MNP are estimated. Due to their excessive numbers, they are not presented in detail nor interpreted. Table 5 shows the correlation of these proba-

[20] In empirical applications, the results of the coefficients—but not necessarily the derived probabilities—are sensitive to the specification of the covariance matrix and exclusion restrictions across choice equations. The empirical identification problem can result in converge problems.

[21] 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 can be used more than once.

TABLE 5.—CORRELATION OF THE ESTIMATED $P^{m|ml}(x)$ OBTAINED FROM THE TEN BINARY PROBIT AND THE MNP

	Basic Training	Further Training	Employment Program	Temporary Wage Subsidy
No participation	0.998	0.989	0.997	0.994
Basic training		0.980	0.991	0.994
Further training			0.992	0.992
Employment program				0.983

Correlations are computed in the sample of participants in the two treatments that define the particular cell.

bilities with those obtained from the MNP in each relevant subsample.

The correlation of the conditional probabilities obtained from the two approaches are indeed very high (between 0.980 and 0.998), so we should expect to obtain basically the same evaluation results irrespective whether the conditional probabilities are derived from the MNP or estimated directly.

C. *Matching Using Different Balancing Scores*

Quality of the Matches: Three variants of matching are implemented as described in table 1. In the following, the term *MNP unconditional* (*MPU*) is used for matching based on both marginal probabilities, *MNP conditional* (*MPC*) denotes the one based on conditional probabilities derived from the MNP, and finally, the matching based on the ten binary probits is termed *binary probit conditional* (*BPC*).

Using the standardized bias as indicator of the match quality, the analysis of the probabilities that are used for matching show that match quality is good in this respect. This indicates that the overlap of these probabilities is generally sufficient.[22] With sufficient support, balancing is implied by the properties of the propensity scores that hold irrelevant of the validity of CIA.

However, the real question is whether matching on these probabilities is sufficient to balance the covariates. Table 6 shows the results for two summary measures—the median absolute standardized bias and the mean squared standardized 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.[23] There is no consensus in the literature regarding how to measure the distance between high-dimensional multivariate distributions with continuous and discrete components, but the two measures given are frequently 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. These measures act as a kind of specification tests for the estimated models, because, if the conditional and the marginal probabilities are correctly specified, balancing of the covariates must be achieved in the absence of a support problem. Thus, the model with lower values is more trustworthy in cases in which the evaluation results from the various approaches differ.

Using the results in table 6 to rank the different versions according to their match quality is difficult. First, comparing the two approaches based on conditional probabilities, it is very hard to spot systematic differences. It seems that all three approaches achieve balancing more or less equally well. This may be seen as indication that the restrictions implied by the MNP formulation are not critical when compared to the reduced form.

A matching algorithm that uses every control group only once runs into problems in regions of the attribute space wherein the density of the probabilities is very low for the control group compared to the treatment group. An algorithm that allows the use of the same observation more than once does not have that problem, as long as there is an overlap in the distributions. The drawback could be that it uses observations too often, in the sense that comparable observations that are almost identical to the ones actually

[22] These results are omitted for the sake of brevity. Similar results can be found in the discussion paper version of this paper, Lechner (2000a), which is downloadable from www.siaw.unisg.ch/lechner. It also contains results for a fourth version of the matching estimator, namely one based only on one marginal probability, ($P^m(x)$). This one, however, appears to be severely biased (as is expected because using only one marginal probability is insufficient to achieve balancing of the covariates).

[23] Again, for the sake of brevity, only the comparison to NO PARTICIPATION and TEMPORARY WAGE SUBSIDY is given in table 6 and the subsequent tables. The entire set of results can be found in the already mentioned discussion paper version of this paper.

TABLE 6.—BALANCING OF COVARIATES: RESULTS FOR THE MEDIAN ABSOLUTE STANDARDIZED BIAS (MASB) AND THE MEAN SQUARED STANDARDIZED BIAS (MSSB)

	MNP Unconditional, $P^m(X)$, $P^l(X)$		MNP Conditional, $P^{m\|ml}$		Binary Conditional, $\tilde{P}^{m\|ml}$		MNP Unconditional, $P^m(X)$, $P^l(X)$		MNP Conditional, $P^{m\|ml}$		Binary Conditional, $\tilde{P}^{m\|ml}$	
l	No Participation						Temporary Wage Subsidy					
m	MASB	MSSB	MASB	MSSB	MASB	MSSB	MASB	MSSB	MASB	MSSB	MASB	MSSB
Basic training	2.8	12	2.7	16	2.5	14	2.6	16	2.9	19	2.1	15
Further training	4.1	29	2.6	22	3.3	24	2.6	21	3.8	15	3.3	18
Employment program	2.3	15	1.9	17	3.3	18	2.9	26	3.8	33	3.7	26
Temporary wage subsidy	2.3	9	2.0	11	2.1	12	—	—	—	—	—	—

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 percent 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 covariates included in the estimation of the MNP. (See table B1.)

TABLE 7.—EXCESS USE OF SINGLE OBSERVATIONS

	MNP Unconditional, $P^m(X)$, $P^l(X)$		MNP Conditional, $P^{m\|ml}$		Binary Conditional, $\tilde{P}^{m\|ml}$		MNP Unconditional, $P^m(X)$, $P^l(X)$		MNP Conditional, $P^{m\|ml}$		Binary Conditional, $\tilde{P}^{m\|ml}$	
l	No Participation						Temporary Wage Subsidy					
m	Top 10	Mean	Top 10	Mean	Top 10	Mean	Top 10	Mean	Top 10	Mean	Top 10	Mean
Basic training	29	1.7	30	1.8	31	1.8	36	2.4	37	2.6	37	2.6
Further training	20	1.2	20	1.3	21	1.3	24	1.5	27	1.6	26	1.6
Employment program	23	1.3	23	1.4	24	1.3	26	1.5	27	1.6	25	1.6
Temporary wage subsidy	24	1.4	25	1.5	24	1.4	—	—	—	—	—	—

Top 10: Share of the sum of largest 10% of weights of total sum of weights. *Mean:* Mean of positive weights.

used are available. Hence, in principle, there could be substantial losses in precision as a price to pay for a reduction of bias.

Table 7 addresses that issue by considering two measures. The first is a concentration ratio that is 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 of the weights for matched comparison observations.

First, it is not a surprising result that both indicators are somewhat higher for the comparison to TEMPORARY WAGE SUBSIDY as to NO PARTICIPATION, because the latter group is larger and contains a wide spread of all probabilities. (See table 4.) Comparing the three estimators, the differences appear to be small, although MPU seems to be somewhat superior in almost all cases (that is, using more observations for the comparison than the other estimator without any loss in terms of insufficient balancing). (See table 6.) Considering tables 6 and 7 together, MPU appears to be somewhat better, although the small differences prohibit any definite judgments.

The Sensitivity of the Evaluation Results with Respect to the Choice of Score: In this section, the issue is the sensitivity of the evaluation results with respect to the choice of propensity scores. Again, to avoid flooding the reader with numbers, table 8 gives the estimation results for the pairwise treatment on the treated effects ($\theta_0^{m,l}$) covering only comparisons of all programs to NO PARTICIPATION and TEMPORARY WAGE SUBSIDY. A positive number indicates that the effect of the program shown in the row on its participants compared to the comparison state given in the respective column is an additional X percentage points of employment. For example, the entry for the fourth treatment ("Temporary w.s.") in first column of the upper panel (MNP unconditional) should be read as "for the population participating in TEMPORARY WAGE SUBSIDY, TEMPORARY WAGE SUBSIDY increases the probability of being employed on day 461 on average by 8.8 percentage points compared to NO PARTICIPATION." Furthermore, results from probit estimation are also added for reference. In the probit estimation, the treatments entered as explanatory variables (four dummy variables) along the explanatory variables used in the MNP estimation of the selection process. (See table B.1.) To ease the comparison of these results to effects such as treatment on the treated, all five mean probabilities corresponding to the different states are computed for each individual and then averaged over the appropriate subpopulation. Then, twenty corresponding differences are formed. In addition, the table also repeats the unadjusted differences for comparison.

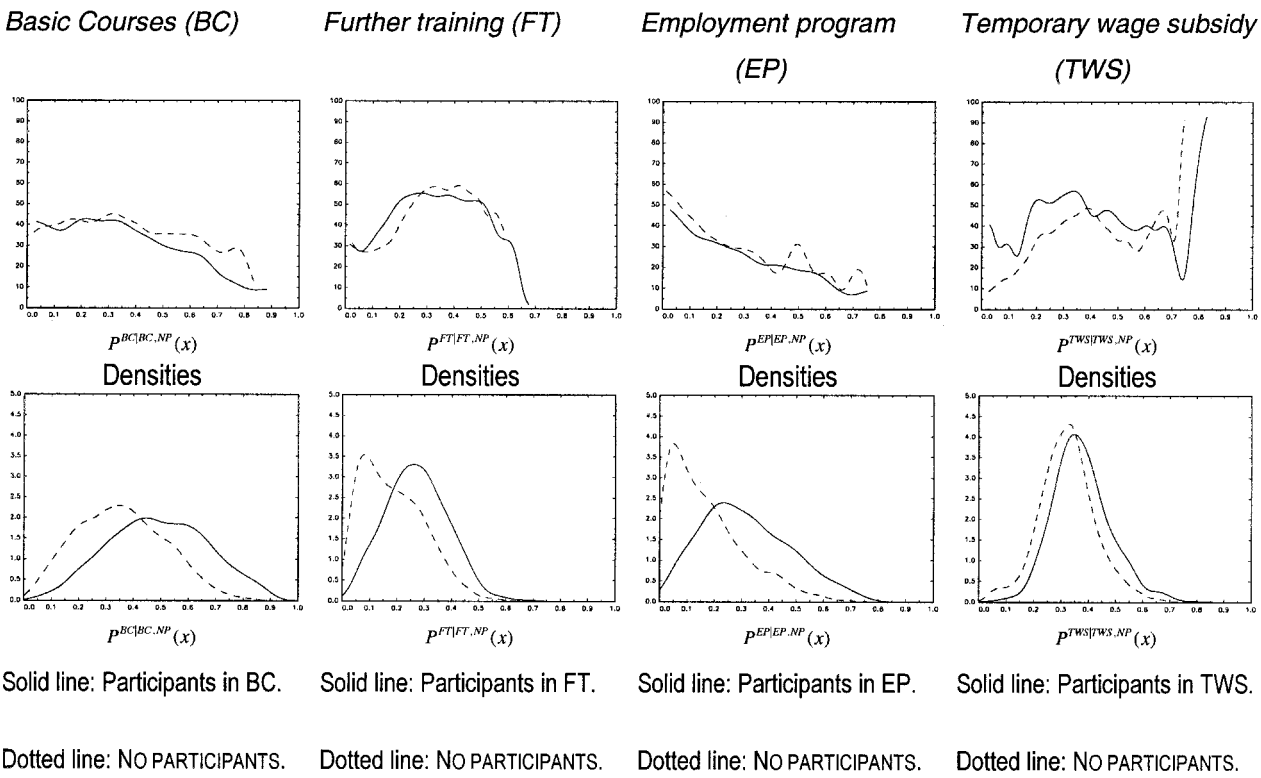
Comparing the three estimators it appears, first of all, that the use of more comparison observations by MPU re-

TABLE 8.—ESTIMATION RESULTS FOR $\theta_0^{m,l}$ IN DIFFERENCES OF PERCENTAGE POINTS

	MNP Unconditional, $P^m(X)$, $P^l(X)$	MNP Conditional, $P^{m\|ml}$	Binary Conditional, $\tilde{P}^{m\|ml}$	Probit Model for Outcomes	Unadjusted Differences
	l: No Participation				
m	Mean (std.)	Mean (std.)	Mean (std.)	Mean	Mean
Basic training	−6.7 (2.4)	−3.1 (2.4)	−6.9 (2.4)	−4.8	−8.6
Further training	.8 (2.9)	1.7 (3.0)	−.7 (2.9)	3.0	10.2
Employment program	−.9 (3.0)	1.3 (3.0)	−6.1 (3.0)	−3.0	−13.0
Temporary wage subsidy	8.8 (2.2)	8.2 (2.2)	8.3 (2.2)	9.0	9.7
m	l: Temporary Wage Subsidy				
Basic training	−13.5 (3.1)	−20.8 (3.0)	−13.7 (3.3)	−14.0	−18.3
Further training	−9.4 (3.2)	−9.9 (3.4)	−10.2 (3.4)	−6.7	0.5
Employment program	−11.0 (3.2)	−15.3 (3.3)	−15.8 (3.3)	−11.8	−22.7

The outcome variable is *employed* for day 461 (in %-points). Standard errors are in brackets.

FIGURE 1.—NONPARAMETRIC REGRESSION OF THE CONDITIONAL PARTICIPATION PROBABILITIES $P^{m|ml}(x)$ ON THE OUTCOME VARIABLE IN RESPECTIVE SUBSAMPLES; COMPARISON STATE: NO PARTICIPATION



Regression: Nadaraya-Watson estimate using a Gaussian kernel and the rule-of-thumb bandwidth. Density: Kernel density estimate using a Gaussian kernel and the rule-of-thumb bandwidth. The results are not very sensitive with respect to bandwidth choice.

sults—as expected—in some cases in slightly smaller (estimated) standard errors. But again the differences are tiny.

Comparing the results column by column, fairly similar conclusions from the three estimators are obtained. Compared to the raw differences, the adjustment always works in the same direction, with one exception. In two of the nine cases, the differences between the largest and the smallest value of the effects are about two standard errors of the single estimate (BASIC TRAINING versus TEMPORARY WAGES SUBSIDY, EMPLOYMENT PROGRAM versus NO PARTICIPATION); in the other cases differences are considerably lower. In the first case, the problem seems to be related to MPC, which balances the covariates worse than the other estimators in that case. (See table 6.) In the second case, BPC appears to be problematic for the same reason. This issue is taken up again when analyzing results of figure 1 in section V.

The first entries in the lower panel of table 8 relate to the probit model for the outcomes. Among other restrictions coming from the functional form of the probit and the linear index specification, it is a major difference compared to the matching approach that the effects are allowed to vary only in a very restrictive way among individuals whereas they can vary freely in the matching approaches.[24] Judged by the range of the results for the matching estimators, the probit seems not to be too bad on average. For the comparison to NO PARTICIPATION, it is however too large (outside the range of the matching results) for BASIC COURSES as well as TEMPORARY WAGE SUBSIDY, as well as for the comparison of EMPLOYMENT PROGRAMS to TEMPORARY WAGE SUBSIDY. In all these cases, the probit estimates are closer to the unadjusted differences than the ones obtained by matching.

V. Heterogeneity of the Effects

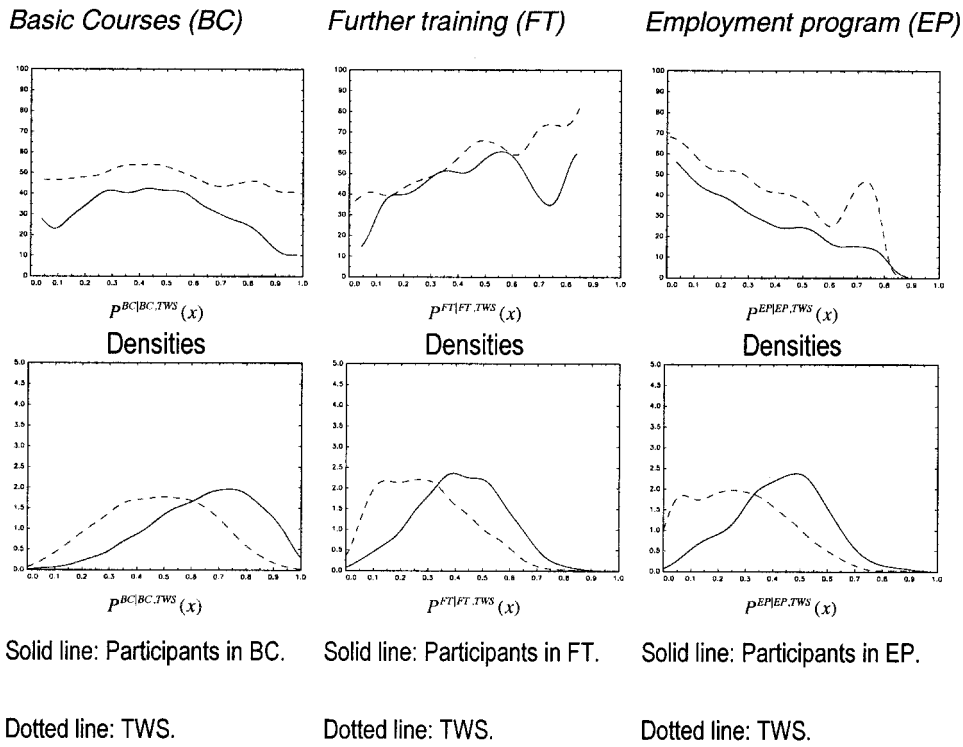
In this section, the issue of heterogeneity of the effects other than by the different types of programs is considered. (The results in this and the following section are all based on MPU.)

A. *Participation Probability*

A question relevant to analyze the efficiency of selection procedures into a program is whether the effects vary with the participation probabilities. Ideally, the effects increase with that probability; that is, the unemployed who are most

[24] Note that, although the coefficients used to parameterize the treatments are the same for all observations, the effects defined in differences of probabilities unconditional on other characteristics vary across subpopulations if the distribution of characteristics vary. The reason is the nonlinearity of the cdf. of the normal distribution.

FIGURE 2.—NONPARAMETRIC REGRESSION OF THE CONDITIONAL PARTICIPATION PROBABILITIES $P^{m|ml}(x)$ ON THE OUTCOME VARIABLE IN RESPECTIVE SUBSAMPLES; COMPARISON STATE: TEMPORARY WAGE SUBSIDY



Regression: Nadaraya-Watson estimate using a Gaussian kernel and the rule-of-thumb bandwidth. Density: Kernel density estimate using a Gaussian kernel and the rule-of-thumb bandwidth. The results are not very sensitive with respect to bandwidth choice.

likely to participate in the programs should benefit most on average. A way to check whether this is true is to consider the expectation of the outcome variable conditional on the conditional selection probabilities ($P^{m|ml}(x)$) in the pool of participants (m) and participants in other states (l). Figure 1 shows such comparisons based on kernel-smoothed regressions for program participants versus NO PARTICIPANTS. Figure 2 presents the same results for the comparison to TEMPORARY WAGE SUBSIDY (TWS). The difference between the curve at any point is an estimate of the causal effect at that specific value of $P^{m|ml}(x)$. Below each nonparametric regression, the smoothed densities of the respective probabilities in the two subsamples are shown because nonparametric regressions are very unreliable in regions of sparse data.

First consider the two programs that already appeared as the ones designated for "bad risks" on the labor market, BASIC COURSE and EMPLOYMENT PROGRAM, in the comparison to NO PARTICIPATION: the employment chances for participants and nonparticipants generally decrease with the participation probability. However, the employment probabilities for the NO PARTICIPANTS are higher (almost) all over the support of the probabilities, and particularly so for high participation probabilities. Hence, we obtain the negative or zero average effects of these programs that appeared before.[25] For EMPLOYMENT PROGRAMS, it seems likely that the difference across estimators spotted in the results of the previous section originate from differently weighting the two little bubbles (regions of negative effects) that appear at high probabilities (particularly the first one carries some weight in the average). For FURTHER TRAINING, the effects are not clear because the regression lines cross twice. It is slightly puzzling that, for higher values of the probabilities (with still enough density), the expected outcome for NO PARTICIPATION is higher. For TEMPORARY WAGE SUBSIDY, the same puzzling feature appears for high probabilities. However, in the region with most of the mass, the regression line for TWS is consistently above the line for NO PARTICIPATION, hence the positive average effect that showed up before. Finally, note that the plots of the densities also suggest that there is no substantial problem of nonoverlapping support, except perhaps for very high probability values for BASIC COURSES and EMPLOYMENT PROGRAMS.

The regression lines of BASIC COURSE and EMPLOYMENT PROGRAM compared to TEMPORARY WAGE SUBSIDY show that TWS dominates unambiguously. The bad news is that the negative effects for BASIC COURSE seem to increase with the

[25] Note that, conceptually, the treatment effect on the treated is a weighted average of the differences of these regression lines, with weights determined by the distribution of the respective participants.

TABLE 9.—ESTIMATION RESULTS FOR $\theta_0^{m,l}$, $\alpha_0^{m,l}$ AND $\gamma_0^{m,l}$

l	No Participation			Temporary Wage Subsidy		
m	$\theta_N^{m,l}$ Mean (std.)	$\alpha_N^{m,l}$ Mean (std.)	$\gamma_N^{m,l}$ Mean (std.)	$\theta_N^{m,l}$ Mean (std.)	$\alpha_N^{m,l}$ Mean (std.)	$\gamma_N^{m,l}$ Mean (std.)
Basic training	−6.7 (2.4)	−4.9 (1.8)	−3.8 (1.8)	−13.5 (3.1)	−12.6 (2.4)	−13.6 (2.1)
Further training	.8 (2.9)	3.3 (3.0)	2.0 (3.3)	−9.4 (3.2)	−4.9 (3.0)	−7.7 (3.5)
Employment program	−.9 (3.0)	−7.8 (3.2)	−6.1 (3.0)	−11.0 (3.2)	−9.9 (2.9)	−15.9 (3.3)
Temporary wage subsidy	8.8 (2.2)	11.0 (2.0)	9.8 (2.0)	—	—	—

The outcome variable is *employed* for day 461 (in percentage points). Standard errors are in brackets.

probability (for larger probabilities), whereas this is not so clear for EMPLOYMENT PROGRAMS. The picture about the effects is not so clear for FURTHER TRAINING because the regressions are fairly close in the center of the distributions and diverge only for regions with fewer observations.

Note that splitting the sample along some characteristics and performing a disaggregated analysis is another possibility to find more subgroup heterogeneity of the effects. However, due to space restrictions, this route is not followed any further in this paper.

B. *Nonparticipants*

When the interest is the effect of the treatment on a person randomly drawn from the population or a person randomly drawn from the participants in that treatment and the comparison treatment, then the treatment effect on the treated is not the correct parameter to analyze. Instead, the following parameters are obvious choices:

$$\gamma_0^{m,l} = E(Y^m - Y^l) = EY^m - EY^l \tag{5}$$

and

$$\begin{aligned}\alpha_0^{m,l} &= E(Y^m - Y^l \big| S \in \{m, l\}) \\ &= E(Y^m \big| S \in \{m, l\}) - E(Y^l \big| S \in \{m, l\}).\end{aligned} \tag{6}$$

$\gamma_0^{m,l}$ denotes the expected (average) effect of treatment m relative to treatment l for a participant drawn randomly from the population. 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 $\alpha_0^{m,l} = -\alpha_0^{l,m}$. Estimation is no more difficult than estimation of $\theta_0^{m,l}$. In fact, only step 4 in table 1 needs to be changed. For details, the reader is referred to Lechner (2001a).

The estimated effects for the different populations are fairly similar. It appears to be surprising that, although the various groups of participants in the different programs are very heterogeneous and although the previous figures suggest that effect heterogeneity is present when defined along the lines of the participation probabilities, the effects for these different populations are not that much different, perhaps with the exception of the comparison of EMPLOYMENT PROGRAMS to NO PARTICIPATION. Such a finding could reinforce the point made in the previous subsection that these programs are not well targeted. The overall conclusion from table 9 is that treatment heterogeneity is important, but population heterogeneity with respect to the effects when defined by groups inside and outside the programs is less so. Because for an efficient selection $\theta_0^{m,l}$ should not be smaller than the other effects, these findings suggest fairly inefficient selection rules into several programs.

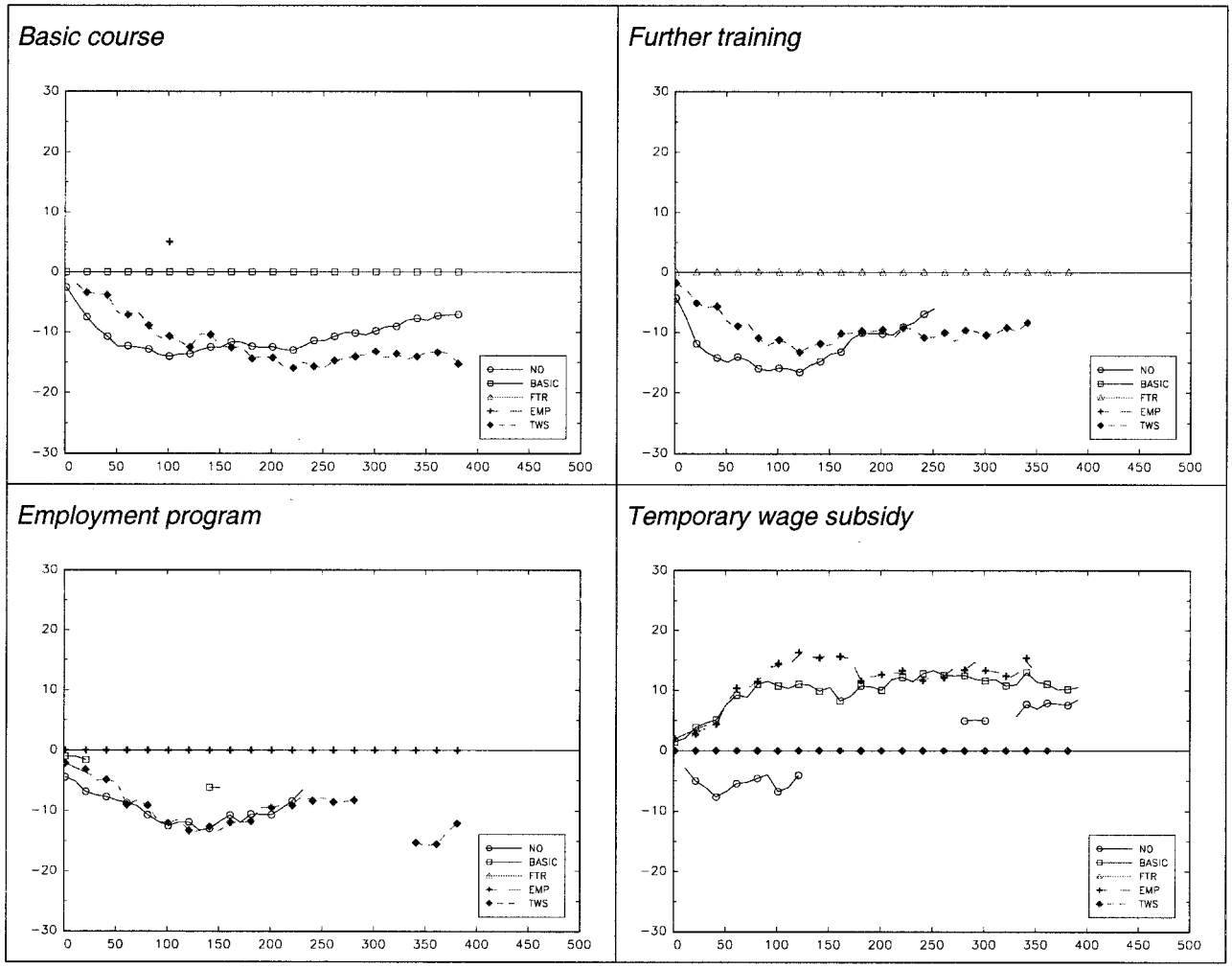
C. *Time*

Finally, the last aspect of heterogeneity considered relates to time. It is conceivable that the differences of the effects between the programs change over time (such as when comparing programs that enhance human capital by training compared to employment programs). Therefore, Figure 3 follows the effects over time from the start of a program (day 1) for all groups of participants compared to all alternative programs. A value larger than zero indicates that the particular program denoted in the header of the figure would increase employment compared to the specific program that is depicted by the particular line.

The results from the perspective given by figure 3 confirms the ordering of the effectiveness of programs established so far. However, in addition, it becomes obvious that the sign as well as the size of the effects depend on the time the effect is measured (*time* is defined as days after the start of the program). From that perspective, it is clear that NO PARTICIPATION should have a positive effect in the short run because the individuals can search for a job more intensively. Thus, it is not surprising that initial negative effects in the comparison to NO PARTICIPATION show up for all programs. However, about one year after the program begins, the initial negative effects are more than overcompensated by the positive effect of TEMPORARY WAGE SUBSIDY. For FURTHER TRAINING and EMPLOYMENT PROGRAM there are positive trends, but the time horizon is perhaps too short for them to materialize. BASIC TRAINING cannot compensate this effect at all and thus appears to be ineffective even one year after its start.

These considerations show that the proposed approach can be used to address the heterogeneity issue in many different ways. Thus, it can become a very useful tool in econometric policy analysis.

FIGURE 3.—EFFECTS OF NONPARTICIPATION COMPARED TO THE PROGRAMS FOR THE POPULATION ($\theta_0^{m,l}$ FOR EMPLOYMENT): TIME RELATIVE TO START OF PROGRAM



Day 1 corresponds to the first day after the start of the program. *NO:* Nonparticipation; *BASIC:* Basic Training; *FTR:* Further training; *EMP:* Employment program; *TWS:* Temporary wage subsidy. A positive number indicates that participation in the respective program increases the employment probability compared to being in one of the other states. The figure displays only mean effects that are significant at the 5% level.

VI. Aggregation

Given the number of treatments of this study, many pairwise comparisons are possible. Hence, a more concise summary measure of the effectiveness of particular programs is useful. To that end, the following composite treatment effects, defined as the weighted sum of the pairwise effects, are introduced:

$$\theta_0^m(\nu^m) = \sum_{l=0}^{M} \nu^{m,l}\theta_0^{m,l};$$

$$\gamma_0^m(\nu^m) = \sum_{l=0}^{M} \nu^{m,l}\gamma_0^{m,l}; \qquad (7)$$

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

$$\nu^{m,m} = 0.$$

Although the composite effects given in equations (7) do not look like causal effects at first sight, they have a causal interpretation if the weights are nonnegative constants that sum to 1: they correspond to the effects of treatment m compared to an artificial state in which 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 as $Y^{-m}(\nu^m) := \sum_{l=0}^{M} \nu^{m,l}Y^l$. Then the composite effects can be rewritten as (see proof in appendix A):

$$\theta_0^m(\nu^m) = E(Y^m|S = m) - E[Y^{-m}(\nu^m)|S = m]. \qquad (8)$$

The same holds for $\gamma_0^m(\nu^m)$ but not for $\alpha_0^m(\nu^m)$. (See appendix A.) Due to the causal interpretation of the composite effects, they could be used to define the effects of treatment m measured relative to some chosen composite alternative treatment.

TABLE 10.—ESTIMATION RESULTS FOR THE COMPOSITE EFFECTS

	$\hat{\theta}_N^m(\tilde{\nu}^m)$	$\hat{\alpha}_N^m(\tilde{\nu}^m)$	$\hat{\gamma}_N^m(\tilde{\nu}^m)$	$\hat{\theta}_N^m(\nu^m)$ (Aggregated)	Unadjusted Differences
Nonparticipation	−1.4	−4.7	−4.4	0.2	0.9
Basic training	−7.1	−10.3	−10.6	−18.2	−10.8
Further training	−0.2	−1.2	−1.8	11.0	11.9
Employment program	−3.0	−5.1	−6.3	−25.1	−13.7
Temporary wage subsidy	9.0	10.1	10.1	27.2	12.7

The outcome variable is *employed* for day 461 (in percentage points). The first three columns are computed from the *MNP unconditional* estimates. The effects presented in the second-to-last column are computed by aggregating the respective nontreatment groups before the estimation of the effect.

The weights specified in the application correspond to the unconditional distribution of treatments other than m in the population excluding participants in m. Although such a specification seems to be intuitive because it weights the other states according to the probability of occurrence, other weighting schemes could also be plausible, depending on the objective of the empirical analysis:

$$\tilde{\nu}^{m,l} = P(S = l \mid S \neq m),$$
$$P(S = l \mid S \neq m) = \frac{P(S = l)}{1 - P(S = m)}, \quad m \neq l. \tag{9}$$

In contrast to $\theta_0^m(\tilde{\nu}^m)$, another parameter $\theta_0^m(\nu^m)$ is introduced and defined by aggregating 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. A probit model is then used to estimate the respective probabilities and an accordingly simplified version (only two categories) of the algorithm outlined in table 1 is used for matching.

Table 10 shows these aggregate effects together with aggregate effects for the two other treatment parameters introduced in the previous section. First, considering the composite effects using $P(S = l \mid l \neq m)$ as weights basically confirms the ranking of the treatments that emerged from the pairwise 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. The latter results in very large values for the estimated effects (in both directions) that do appear to be plausible at all.

Note that $\theta_0^m(\nu^m)$ can easily be computed because it is based on the simple binary treatment model. Therefore, from a practical point of view, an interesting question arises: Does $\theta_0^m(\nu^m)$ like $\theta_0^m(\tilde{\nu}^m)$ correspond to a particular weighting scheme and thus have a causal interpretation? The answer is yes, it has a causal interpretation, but it is difficult to derive the weights (ν^m) explicitly, because they depend on the particular distribution of $P^m(x)$ in the specific comparison groups:

$$\begin{aligned}\theta_0^m(\nu^m) &= E(Y^m \mid S = m) - E(Y^{-m} \mid S = m) = \\ &= E(Y^m \mid S = m) - \\ &\quad - E_X\{E[Y^{-m} \mid P^m(X), S = -m] \mid S = m\}.\end{aligned} \tag{10}$$

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 $\theta_0^m(\tilde{\nu}^m)$ and $\theta_0^m(\nu^m)$ are in general different causal effects. Because the latter is difficult to express explicitly, it has no interpretable meaning in economic terms. Thus, explicitly aggregating effects by using composite effects like $\theta_0^m(\tilde{\nu}^m)$ seems to be a useful approach to condense the information from the pairwise effects, whereas aggregating heterogeneous groups of participants $(\theta_0^m(\nu^m))$ can lead to fairly misleading conclusions.

VII. Conclusion

This paper suggests an approach of handling the issue of treatment heterogeneity in microeconometric evaluation studies based on propensity score matching. The proposed methods are applied to the evaluation of different programs of Swiss active labor market policies to provide an example for their potential use.

The paper addresses the issues of individual heterogeneity and treatment heterogeneity and shows that the multiple-treatment approach can lead to valuable insights. The paper also proposes summary measures of causal effects for different treatments and discusses their causal interpretation. It shows that an effect based on a comparison of a treatment group to an aggregated comparison group of individuals has no meaningful causal interpretation and can lead to misleading results. However, appropriately aggregated pairwise effects give a clear-cut causal effect that could be effectively used to rank different programs.

Different approaches to modeling the respective propensity scores needed for matching are also discussed. One approach consists of deriving the probabilities used for the propensity scores by specifying and estimating a multiple discrete-choice model, such as a multinomial probit model. The alternative is to concentrate on modeling and estimating all conditional probabilities between possible pairs of choices directly. One advantage of using a multinomial discrete-choice model instead of concentrating only on binary conditional choices is that it is easier to understand the empirical factors behind the joint selection process. The drawback is that it is computationally 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 particular three matching estimators suggested for the multiple-treatment framework give roughly the same answers.

REFERENCES

Angrist, J. D., "Estimating Labor Market Impact of Voluntary Military Service Using Social Security Data," *Econometrica* 66:2 (1998), 249–288.

Angrist, J. D., and A. B. Krueger, "Empirical Strategies in Labor Economics" (pp. 1277–1366), in O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics,* vol. III A, (Amsterdam: North-Holland, 1999).

Brodaty, Th., B. Crepon, and D. Fougère, "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986–1988" (pp. 85–123), in M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluations of Active Labor Market Policies in Europe* (Heidelberg: Physica/Springer, 2001).

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

Dehejia, R. H., and S. Wahba, "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94:448 (1999), 1053–1062.

Frölich, M., "Treatment Evaluation: Matching versus Local Polynomial Regression," University of St. Gallen discussion paper no. 2000-17 (2000).

Gerfin, M., and M. Lechner, "A Microeconometric Evaluation of the Active Labor Market Policy in Switzerland," University of St. Gallen discussion paper no. 2000-08 (2000).

Hajivassiliou, V. A., and P. A. Ruud, "Classical Estimation Methods for LDV Models Using Simulation" (pp. 2384–2441), in R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics,* vol. IV (Amsterdam: North-Holland, 1994).

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

Heckman, J. J., R. J. LaLonde, and J. A. Smith, "The Economics and Econometrics of Active Labor Market Programs" (pp. 1865–2097), in O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics,* vol. III A (Amsterdam: North-Holland, 1999).

Holland, P. W., "Statistics and Causal Inference," *Journal of the American Statistical Association* 81:396 (1986), 945–970, with discussion.

Imbens, G. W., "The Role of the Propensity Score in Estimating Dose-Response Functions," NBER technical working paper no. 237 (1999). Also in *Biometrika* 87 (2000), 706–710.

Keane, M. P., "A Note on Identification in the Multinomial Probit Model," *Journal of Business & Economic Statistics* 10:2 (1992), 193–200.

Larsson, L., "Evaluation of Swedish Youth Labour Market Programmes," Office for Labour Market Policy Evaluation (Uppsala) discussion paper no. 2000:1 (2000).

Lechner, M., "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification," *Journal of Business & Economic Statistics* 17:1 (1999), 74–90.

———, "Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies," University of St. Gallen discussion paper no. 2000-01 (2000a).

———, "Some Practical Issues in the Evaluation of Heterogeneous Labor Market Programs by Matching Methods," University of St. Gallen discussion paper no. 2000-14 (2000b).

———, "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption" (pp. 43–58), in M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluations of Active Labor Market Policies in Europe* (Heidelberg: Physica/Springer, 2001a).

———, "A Note on the Common Support Problem in Applied Evaluation Studies," University of St. Gallen discussion paper no. 2001-01 (2001b).

McFadden, D., "Econometric Models of Probabilistic Choice," in C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: The MIT Press, 1981).

———, "Econometric Analysis of Qualitative Response Models" (pp. 1396–1457), in Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics,* vol. 2 (Amsterdam: North-Holland, 1984).

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

———, "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79:387 (1985), 516–524.

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

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

———, "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics* 2 (1977), 1–26.

APPENDIX A: TECHNICAL APPENDIX

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

$$\begin{aligned}
\theta_0^m(\nu^m) &= \sum_{l=0}^{M} \nu^{m,l}\theta_0^{m,l} = \sum_{l=0}^{M} \nu^{m,l}[E(Y^m|S=m) - E(Y^l|S=m)] \\
&= E(Y^m|S=m) - \sum_{l=0}^{M} \nu^{m,l} E(Y^l|S=m) \\
&= E(Y^m|S=m) - E\left[\left(\sum_{l=0}^{M} \nu^{m,l} Y^l\right)\middle| S=m\right] \\
&= E(Y^m|S=m) - E[Y^{-m}(\nu^m)|S=m]. \qquad \text{q.e.d.}
\end{aligned}$$

The same line of argument is valid for $\gamma_0^m(\nu^m)$ as well:

$$\begin{aligned}
\gamma_0^m(\nu^m) &= \sum_{l=0}^{M} \nu^{m,l}\gamma_0^{m,l} = \sum_{l=0}^{M} \nu^{m,l}[E(Y^m) - E(Y^l)] \\
&= E\left(Y^m - \sum_{l=0}^{M} \nu^{m,l} E(Y^l)\right) \\
&= E(Y^m) - E\left(\sum_{l=0}^{M} \nu^{m,l} Y^l\right) \\
&= E(Y^m) - E[Y^{-m}(\nu^m)]. \qquad \text{q.e.d.}
\end{aligned}$$

Furthermore, note that such an interpretation does not appear to be available for $\alpha_0^m(\nu^m)$:

$$\begin{aligned}
\alpha_0^m(\nu^m) &= \sum_{l=0}^{M} \nu^{m,l}\alpha_0^{m,l} \\
&= \sum_{l=0}^{M} \nu^{m,l}[E(Y^m|S\in\{m,\,l\}) - E(Y^l|S\in\{m,\,l\})].
\end{aligned}$$

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

APPENDIX B: ESTIMATION OF MULTINOMIAL PROBIT MODEL

The multinomial probit model with more than four categories is computationally untractable because the choice probabilities are high-dimensional integrals without a closed-form analytical representation. Among others, Börsch-Supan and Hajivassiliou (1993) and Hajivassiliou and Ruud (1994) show, however, that the MNP maximum likelihood estimator can nevertheless be approximated by simulation methods. Drawing on their results about accuracy and simulation bias of the estimates, the MNP is estimated by simulated maximum likelihood using the GHK simulator with four hundred independent draws for each individual and choice equation. Given the available Monte Carlo evidence, four hundred draws should result in an almost negligible simulation error. In fact, varying the number of draws does lead to only minor changes in the estimated coefficients.

The covariance matrix of the MNP error terms is not fully identified, so normalizing constraints need to be imposed. (See, for example, Keane (1992).) Furthermore, they are necessary to avoid excessive numerical instability in finite samples. It can be seen from the lower part of table B1 that some of these restrictions are imposed, basically concerning the variances and the correlations of the reference groups with the other groups. Although the MNP is in principle identified without further restrictions on the variables of the choice equations (other than normalizing the coefficient of the reference group to zero), in practice such exclusion restrictions seem to drive the result. Therefore, they are minimized and related to the institutional setting.[26] As can be seen from the upper part of table B1 the information about the mother tongue is fully allowed only in the equation relating to BASIC COURSES (consisting of a large share of language courses). Furthermore, some specifics about relevant sectors and occupations are also excluded from some choice equations. All other variables appear in all equations.

The correlations between the choice specific error terms vary between −0.9 and 0.4. The high negative correlation as well as the general lack of precision of the covariance matrix estimate is a somewhat worrying feature and may point to an identification problem. The lack of precision is transferred to the other estimated coefficients, particularly those of the equation relating to further training (for which the variance of the error is free). A Wald test cannot reject the null of zero correlations. (See note on table B1.) Sensitivity checks with some more-restrictive as well as some more-general covariance structures reveal that the evaluation results (that depend only on probabilities and not on coefficients directly) hardly change. Within the MNP, however, in the actual specification there appears to be a considerable increase in the estimated standard errors of the two groups with the largest estimated variances—FURTHER TRAINING and TEMPORARY WAGE SUBSIDY—compared to some more-restrictive specifications.

[26] Entries for variables excluded from a particular choice equation show a zero for the coefficient and "−" for the standard error. Sensitivity checks with respect to some exclusion restrictions (particularly with respect to the variable *duration of previous unemployment spell*) do not indicate much sensitivity.

TABLE B1.—RESULTS OF THE ESTIMATION OF A MULTINOMIAL PROBIT MODEL

	Basic Training		Further Training		Employment Program		Temporary Wage Subsidy	
	coef.	std.	coef.	std.	coef.	std.	coef.	std.
Constant	**1.31**	0.32	−3.88	3.45	*−1.91*	1.01	−0.84	0.85
Age in years/10	0.02	0.03	0.21	0.16	**0.17**	0.06	0.07	0.05
Gender: female	**0.35**	0.05	−0.60	0.49	*−0.24*	0.14	−0.03	0.10
Married	0.03	0.06	−0.37	0.27	**−0.43**	0.15	−0.11	0.09
First foreign language:								
French, Italian, German	**−0.30**	0.06	−0.07	0.21	*0.28*	0.13	0.14	0.10
Native language:								
French	**0.72**	0.18	0	—	0	—	0	—
Italian	**0.69**	0.10	0	—	0	—	0	—
Other than French, Italian, German	**0.76**	0.08	−0.78	0.56	−0.17	0.13	−0.26	0.17
Temporary foreign resident (work permit B)	**0.32**	0.07	−1.28	1.06	0.09	0.12	−0.12	0.15
Information about local labor office								
located in labor market region: small villages	**−1.86**	0.33	−0.22	0.76	*−0.75*	0.42	0.10	0.46
located in labor market region: big cities	**−0.74**	0.12	−1.16	0.81	**−0.65**	0.20	−0.16	0.17
share of entry into long-term unemployed of all UE	**−0.65**	0.19	0.41	0.64	*0.72*	0.38	−0.02	0.28
no information on shares available	**−1.62**	0.26	0.06	0.82	0.26	0.45	0.21	0.44
Subjective valuations of labor office qualification:								
best chance to find a new job:	**−0.19**	0.05	1.07	0.82	0.001	0.09	0.09	0.11
unclear	0.07	0.09	−0.67	0.57	**−0.53**	0.21	0.13	0.16
very easy	−0.07	0.19	−0.07	0.53	−0.53	0.35	0.14	0.27
easy	−0.01	0.08	−0.09	0.25	−0.10	0.13	*0.25*	0.15
difficult	0.16	0.06	−1.01	0.75	−0.06	0.11	*−0.30*	0.16
special case	−0.32	0.16	−1.19	0.86	−0.15	0.22	**−1.58**	0.64
Desired level of employment: part-time	**−0.34**	0.07	0.001	0.27	**−0.55**	0.14	**−0.36**	0.13
Last sector								
construction	−0.09	0.08	−0.68	0.57	**−0.37**	0.15	0.25	0.18
public services	**0.41**	0.09	−0.26	0.36	−0.01	0.15	−0.46	0.28
communications, news	0.52	0.38	1.40	1.35	0.33	0.48	0.66	0.55
tourism, catering	*0.17*	0.07	−0.99	0.78	−0.20	0.13	−0.01	0.12
services (properties, renting, leasing, . . .)	0	—	0	—	*−1.25*	0.58	0	—
other services	0	—	0	—	0	—	*0.55*	0.26
Last occupation								
transportation	*−0.32*	0.15	0	—	0	—	0	—
office	0	—	1.12	0.86	0	—	0	—
architects, engineers, technicians	0	—	2.01	1.57	0	—	0	—
education	0	—	0	—	0	—	*0.88*	0.42

TABLE B1.—*(CONTINUED)*

	Basic Training		Further Training		Employment Program		Temporary Wage Subsidy	
	coef.	std.	coef.	std.	coef.	std.	coef.	std.
Desired occupation same as last occupation	**−0.13**	0.05	−0.16	0.19	−0.11	0.07	0.03	0.08
Previous job position: high (management, . . .)	−0.01	0.11	0.51	0.58	**−0.64**	0.23	*−0.32*	0.19
Duration of previous unemployment spell/1000	**−0.91**	0.15	−0.98	0.93	0.47	0.31	−0.39	0.26
Duration of CUES until start of program/1000	**−2.47**	0.37	−0.12	1.25	−0.33	0.48	**−2.29**	0.81
Duration of CUES								
less than 90 days	−0.09	0.08	−0.54	0.44	**−0.69**	0.27	−0.02	0.12
less than 180 days	−0.05	0.08	0.31	0.42	**−0.55**	0.20	*−0.27*	0.15
Days from 12/31/97 until start/100	*0.09*	0.04	−0.12	0.28	**0.43**	0.10	**0.41**	0.14

	No Participation		Basic Training		Further Training		Employment Program		Temporary Wage Subsidy	
	Implied Covariance Matrix of the Error Terms*									
	coef.	*t*-val.	coef.	*t*-val.	coef.	*t*-val.	coef.	*t*-val.	coef.	*t*-val.
No participation	1	—	0	—	0	—	0	—	0	—
Basic training			1	—	−3.6	−1.2	−0.6	−1.2	−0.6	−1.0
Further training					13.7	—	1.9	−0.03	−1.3	−0.5
Employment program							1.5	—	−0.6	−0.9
Temporary wage subsidy							—		3.0	—
	Implied Correlation Matrix of the Error Terms									
No participation	1		0		0		0		0	
Basic training			1		−0.96		−0.51		−0.32	
Further training					1		0.42		0.20	
Employment program							1		−0.28	

Simulated maximum likelihood estimates using the GHK simulator (four hundred draws of simulator for each observation and choice equation). Coefficients of the category NO PARTICIPATION 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: −10262.8.

Bold numbers indicate significance at the 1% level (two-sided test); numbers in *italics* relate to the 5% level.

If not stated otherwise, all information in the variables relates to the last day of December 1997.

* Six 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).

Wald test for noncorrelated error terms: $\chi^2(6) = 4, 1$ (p-val: 66%).