Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 125 (2005) 15–51
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs

Arild Aakvik[a,1], James J. Heckman[b,c,d,*,2], Edward J. Vytlacil[e,2]

[a] *Department of Economics, University of Bergen, Fosswinckelsgt. 6, N-5007 Bergen, Norway*
[b] *Department of Economics, the University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA*
[c] *American Bar Foundation, 750 North Lake Shore Drive, Chicago, IL 60611, USA*
[d] *Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom*
[e] *Department of Economics, Stanford University, Landau Economics Building, Room 231, 579 Serra Mall, Stanford, CA 94305, USA*

Available online 28 July 2004

Abstract

This paper analyzes the impact of interventions on discrete outcomes when responses to treatment vary among observationally identical persons. Using a latent variable model motivated by economics, we show how to define and identify various mean treatment effects as well as the distribution of treatment effects for discrete outcomes. The framework is based on discrete choice models with unobservables generated by factor structures. Responses to treatment vary among persons who are observationally identical, and agents participate in the program on the basis of their idiosyncratic response to treatment. We apply the model to study the Norwegian Vocational Rehabilitation training program.

JEL classification: C50; H43; J24; J64

Keywords: Social program evaluation; Discrete-choice models; Vocational rehabilitation

*Corresponding author. Tel.: +1-773-702-0634; fax: +1-773-702-8490.
E-mail addresses: arild.aakvik@econ.uib.no (A. Aakvik), jjh@uchicago.edu (J.J. Heckman), vytlacil@stanford.edu (E.J. Vytlacil).

[1] Supported by the University of Bergen, the Meltzer Foundation, and the Foundation for Research in Economics and Business Administration.
[2] Supported by NSF 97-09-873 and NIH:R01-HD34958-01 and grants from the Mellon, Spencer and Donner Foundations.

0304-4076/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2004.04.002

1. Introduction

This paper formulates and estimates an econometric model for evaluating social programs when outcomes are discrete and responses to treatment vary among observationally identical persons. The model can be used to generate a variety of mean treatment effects (treatment on the treated, the average treatment effect and the marginal treatment effect) from a common set of parameters as well as distributions of treatment effects defined on various subpopulations. The latent variables that generate our model can be used to capture the essential features of a variety of well-posed economic models and allow us to bridge the literatures on structural estimation and program evaluation. Estimates produced from our model are economically interpretable and can be used to conduct out-of-sample forecasts and to pool evidence across studies—the usual benefits of a structural econometric approach.

Discrete outcomes arise in analyses of employment, health and migration. Yet a substantial amount of research in the evaluation literature assumes outcomes are continuous or makes special assumptions for analysis of discrete data outcomes. (see, e.g., Card and Sullivan, 1988; Gritz, 1993; Gay and Borus, 1980; Heckman and Robb, 1985; Ham and Lalonde, 1996; Ridder, 1986). The methodology we propose and implement in this paper and a companion technical paper (Aakvik et al., 1999) is sufficiently flexible to accommodate discrete, continuous and mixed discrete-continuous outcome (e.g. Tobit) variables and can be generalized to panel data settings. (see Carneiro et al., 2003). In this paper, we focus on single-period models with discrete outcomes.

We apply our methods to estimate the impact of Norwegian Vocational Rehabilitation Programs (VR) on employment for female applicants. These programs offer income maintenance payments and training programs to individuals whose medical conditions result in reduced productivity in the labor market and whom program administrators believe will benefit from these services. The primary goal of these programs is to allow recently disabled persons to reenter the labor force.

We use this application to illustrate how our methodology can address four questions: (1) What is the overall effect of training on employment probabilities? (2) Which groups of individuals benefit most from participation in training? (3) How important is it to control for unobservables in understanding the selection and outcome processes? (4) What are the effects of applicants' observed and unobserved characteristics on the administrative decision to accept applicants into an on-going training program?

At first glance, the Norwegian VR training program appears to be successful. Those women who receive training have employment rates that are higher than those who do not receive training. However, our results suggest that the apparent success of the program is due to selection—controlling for selection on observable and unobservable characteristics results in a negative estimated average effect from the training. The results suggest that training helps some individuals, particularly those with observed and unobserved characteristics that make them the least likely to return to the labor force without training. However, program administrators only infrequently select such individuals into training.

Our empirical estimates are very imprecisely estimated. Our estimates are at best suggestive because the standard errors are such that we cannot reject the null hypothesis of no selection on unobservables. We use alternative methods to control for selection on unobserved characteristics into training to test the sensitivity of our estimates to alternative identifying assumptions. This analysis produces a range of estimated treatment effects, but all such estimated treatment effects are lower than those estimated effects that only control for observed characteristics. Furthermore, all methods that allow for selection produce estimates of the effect of treatment on the treated that are lower than the corresponding estimates for the average treatment effect.

Another contribution of this paper is to the definition and identification of cream-skimming on observables and unobservables as perceived by the observing economist. A variety of definitions of cream-skimming exist in the literature (see Heckman et al., 2002). For example, in some studies cream-skimming is said to occur when program administrators systematically admit persons who would likely have high employment rates and earnings even in the absence of the program (see, e.g., Bassi, 1983; Anderson et al., 1993). We present several rigorous definitions of the concept and present methods for determining the empirical importance of cream-skimming on both observed and unobserved characteristics. We find substantial evidence for *perverse* cream-skimming. People selected into the VR program have both observable and unobservable factors that produce the lowest *gains* in employment compared to what they would experience without treatment.

This paper is organized in the following way. In Section 2, we present a class of latent variable models that can be used to generate and produce structure on the Neyman (1923), Fisher (1935), Cox (1958) and Rubin (1974) model of potential outcomes, that can be used to estimate structural econometric models and that can be used to analyze discrete, continuous and mixed discrete-continuous outcomes.[3] In Section 3, we define commonly used treatment effect parameters in terms of the latent variables, using, as a unifying device, the marginal treatment effect parameter (*MTE*) introduced in Björklund and Moffit (1987) and Heckman (1997). We consider both means and distributions of treatment effects. In Section 4 we present a factor structure model and in Section 5 we discuss empirical implementation and estimation of the model. Section 6 presents background on the Norwegian VR training program and discusses the data. In Section 7, we present estimates of the model that illustrate the potential of our method. The paper concludes with a summary in Section 8.

2. Latent variable model

In this paper, we use the latent variable model of Heckman and Vytlacil (1999). For each person i, assume two potential outcomes (Y_{0i}, Y_{1i}) corresponding, respectively, to the potential outcomes in the untreated and treated states. Multiple

[3] Heckman and Vytlacil (2005) relate these statistical models to the causal models of economics.

outcome models are analyzed in Aakvik et al. (1999) Carneiro, Hansen and Heckman (2003) and Heckman and Vytlacil (2005). Let $D_i = 1$ denote the receipt of treatment; $D_i = 0$ denotes nonreceipt. Let Y_i be the measured outcome variable so that

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

This is the Neyman–Fisher–Cox–Rubin model of potential outcomes. It is also the switching regression model of Quandt (1972) or the Roy model of income distribution (Roy, 1951; Heckman and Honoré, 1990).

The potential outcome equation for the participation state is

$$Y_{1i} = \mu_1(X_i, U_{1i}), \tag{1}$$

and the potential outcome for the non-participation state is

$$Y_{0i} = \mu_0(X_i, U_{0i}), \tag{2}$$

where X_i is a vector of observed random variables and (U_{1i}, U_{0i}) are unobserved random variables. It is assumed that Y_{0i} and Y_{1i} are defined for everyone and that these outcomes are independent across persons so that there are no interactions among agents. We assume that the program being evaluated is small so that general equilibrium effects and social interactions are negligible.[4]

This paper assumes that a latent variable model generates the indicator variable D. Specifically, we assume that the assignment or decision rule for the indicator is generated by a latent variable D_i^*:

$$\begin{aligned} &D_i^* = Z_i\beta_D - U_{Di} \\ &D_i = 1 \text{ if } D_i^* \geqslant 0, \ D_i = 0 \text{ otherwise,} \end{aligned} \tag{3}$$

where Z_i is a vector of observed (by the econometrician) random variables and U_{Di} is an unobserved (by the econometrician) random variable. D_i^* is the net utility or gain to the decision-maker from choosing state 1. The index structure underlies many models in econometrics (see, e.g., Maddala, 1983) and in psychometrics (see, e.g., Junker and Ellis, 1997).[5] We assume access to an i.i.d. sample, and will henceforth suppress the i subscripts.

In our empirical application, the outcome variable is dichotomous, and we assume that a linear latent index generates the outcome:

$$\mu_j(X, U_j) = \mathbf{1}[X\beta_j \geqslant U_j], \tag{4}$$

[4] Heckman et al. (1998b, c) demonstrate the dangers of ignoring these interactions for large-scale programs. Lewis (1963) discusses this problem in the context of estimating union relative wage effects. The program we study in our application is the Norwegian Vocational Rehabilitation program. This program directly affects 1.5% of the working age population at any given time. While large for a vocational rehabilitation program, we do not believe the program to be large enough to make general equilibrium effects a first-order concern.

[5] Vytlacil (2002) shows that the non-parametric version of the latent index model considered in this paper is equivalent to the assumptions imposed in the local average treatment effect framework of Imbens and Angrist (1994).

where $j = 1$ for the treated state and $j = 0$ for the non-treated state, and where $\mathbf{1}[\cdot]$ is the indicator function. However our methods apply more generally to the cases where (Y_{0i}, Y_{1i}) are discrete, continuous or mixed discrete-continuous. (see Aakvik et al., 1999 and Carneiro, Hansen and Heckman 2003). We can also analyze more general choice processes using the analysis of Matzkin (1992, 1994).

Throughout most of this paper we will maintain the following assumptions:

(i) $Z\beta_D$ is a non-degenerate random variable conditional on $X = x$.
(ii) (U_D, U_1) and (U_D, U_0) are absolutely continuous with respect to Lebesgue measure on $\Re^2$.
(iii) (U_D, U_1) and (U_D, U_0) are independent of (Z, X).
(iv) Y_1 and Y_0 have finite first moments.
(v) $1 > \Pr(D = 1 \mid X) > 0$.

Assumption (i) requires an exclusion restriction: there exists a variable that determines the treatment decision but does not directly affect the outcome. Assumption (ii) is made for technical convenience and can be relaxed. Assumption (iii) can easily be weakened to make (U_D, U_1) and (U_D, U_0) independent of Z given X. This is a standard instrumental variable assumption. Assumption (iv) is required if mean treatment parameters are to be well defined, and is satisfied trivially when Y_1 and Y_0 are binary. Assumption (v) is the standard assumption that for each set of X variables, we observe people in both treated and untreated states, at least in large samples.

Let F_{U_D} be the distribution of U_D, with the analogous notation for the distribution of the other random variables. We define the joint distributions for the unobservables using the notation $F_{D,1} = F_{U_D,U_1}$ and $F_{D,0} = F_{U_D,U_0}$.

3. Treatment parameters

An important advantage of the latent variable model developed in this paper is that it can be used to generate mean treatment parameters and distributions of treatment parameters from a common set of structural parameters. In Section 3.1, we analyze the mean treatment parameters following the analysis of Heckman and Vytlacil (1999). In Section 3.2, we extend the analysis of Heckman and Vytlacil (1999) by considering distributional treatment parameters instead of mean treatment parameters. The analysis in Sections 3.1 and 3.2 extends the previous literature and provides the basis for the analysis of factor structure models in Section 4.

3.1. Mean treatment parameters

Let Δ denote the treatment effect for a given observation, where $\Delta = Y_1 - Y_0$. This person-specific treatment effect is a counterfactual. For a given individual, what would be his or her outcome if he or she received the treatment compared to the case where the person had not received the treatment? One can rarely estimate Δ for any

person.[6] Instead, it is more common to work with population means or distributions of these variables. In this section, we examine three different mean parameters within this framework: the marginal treatment effect (*MTE*), the average treatment effect (*ATE*), and the effect of treatment on the treated (*TT*). We consider the distributional parameters corresponding to each mean parameter in the next section. Each mean parameter corresponds to an average value of Δ but defined on different conditioning sets. *MTE* gives the average effect for persons who are indifferent between participating or not for a given value of the instrument. *ATE* is the average effect for an individual chosen at random from the population of training applicants. *TT* is the average effect for persons who participate. We consider estimation of the distributions of potential outcomes in the next section. We first define the treatment parameters more generally, and then specialize to the case where the outcome variable is generated by a latent index model.

The first parameter we consider is the marginal treatment effect (*MTE*) parameter introduced in Heckman (1997).[7] Following Heckman (1997), we define the *MTE* parameter as

$$\Delta^{MTE}(x,u) = \mathrm{E}(\Delta|X=x, U_D=u). \tag{5}$$

$\Delta^{MTE}(x,u)$ is the average effect of participating in the program for people who are on the margin of indifference between participation in the program ($D=1$) or not ($D=0$) if the instrument is externally set so that $Z\beta_D = u$. For small values of u, $\Delta^{MTE}(x,u)$ is the average effect for individuals with unobserved characteristics that make them the most inclined to participate in the program ($D=1$), and for large values of u it is the average treatment effect for individuals with unobserved (by the econometrician) characteristics that make them the least inclined to participate. High u is associated with high net cost.

The second parameter that we consider in this section is the average effect of treatment on a person selected randomly from the population of individuals with a given value of X. The average treatment effect is given by

$$\Delta^{ATE}(x) \equiv \mathrm{E}(\Delta|X=x).$$

This is related to the marginal treatment effect via the following equation:

$$\Delta^{ATE}(x) = \int \mathrm{E}(\Delta|X=x, U_d=u)\,\mathrm{d}F(u),$$

where integration is made over the full support of U_D. *ATE* is an average of the *MTE* parameters. An average treatment effect integrated over the distribution of X is often desired to summarize data. Thus the following parameter is sometimes sought:

$$\mathrm{E}(\Delta^{ATE}) = \int \Delta^{ATE}(x)\,\mathrm{d}F_X(x).$$

[6] Some panel data estimators identify Δ for each person. See the discussion in Heckman and Smith (1998) and Heckman et al. (1999).

[7] See also Heckman and Smith (1998) and Heckman and Vytlacil (1999, 2000, 2001). A version of this parameter was introduced in a generalized Roy model by Björklund and Moffit (1987). It can also be viewed as the limit form of the LATE parameter of Imbens and Angrist (1994) and Angrist et al. (2000).

Averages over subsets of the support of X are also sometimes of interest. The marginal effect of changes in X on the average treatment effect integrating over the distribution of X is sometimes of interest. Let x_k denote the kth element of X, and assume that $\mathrm{E}(\mu_j(X, U_j)|X = x)$ is differentiable in x_k a.e. F_X for $j = 0, 1$, then a parameter of interest is

$$\mathrm{E}_X\left(\frac{\partial \mathrm{E}(\Delta|X = x)}{\partial x_k}\right) = \int \frac{\partial \Delta^{ATE}(x)}{\partial x_k}\,\mathrm{d}F_X(x).$$

The mean effect of treatment on the treated is the most commonly estimated parameter,[8] and in our notation it is

$$\begin{aligned}\Delta^{TT}(x, D = 1) &\equiv \mathrm{E}(\Delta|X = x, D = 1),\\ &= \mathrm{E}(\Delta|X = x, U_D \leqslant Z\beta_D).\end{aligned} \tag{6}$$

This parameter is the effect of treatment on an individual drawn at random from the population of individuals who entered the program and have the given value of X. The average marginal effect of changes in X on the effect of treatment on the treated is sometimes sought and can be obtained by integrating over the distribution of X for participants:

$$\mathrm{E}_X\left(\frac{\partial \mathrm{E}(\Delta|D = 1, X = x)}{\partial x_k}\middle| D = 1\right) = \int \frac{\partial \Delta^{TT}(x, D = 1)}{\partial x_k}\,\mathrm{d}F_{X|D}(x|1).$$

It will be useful to define a conditional on Z version of $\Delta^{TT}(x, D = 1)$, where Z denotes the regressors that directly enter the selection rule:

$$\begin{aligned}\Delta^{TT}(x, z, D = 1) &\equiv \mathrm{E}(\Delta|X = x, Z = z, D = 1)\\ &= \mathrm{E}(\Delta|X = x, U_D \leqslant z\beta_D)\\ &= \frac{1}{\Pr[D = 1|Z = z]}\int_{-\infty}^{z\beta_D} \mathrm{E}(\Delta|X = x, U_D = u)\,\mathrm{d}F_U(u).\end{aligned}$$

The two versions of TT are related by the following expression:

$$\Delta^{TT}(x, D = 1) = \mathrm{E}(\Delta|X = x, D = 1) = \int \Delta^{TT}(x, z, D = 1)\,\mathrm{d}F_{Z|X,D}(z|x, 1).$$

Using Bayes' rule and the fact that $\Pr(D = 1|X = x, Z = z) = \Pr(D = 1|Z = z)$, we obtain

$$\mathrm{d}F_{Z|X,D}(z|x, 1) = \frac{\Pr(D = 1|Z = z)}{\Pr(D = 1|X = x)}\,\mathrm{d}F_{Z|X}(z|x), \tag{7}$$

[8] See Heckman and Robb (1985) and Heckman et al. (1999).

so that we can obtain an expression in terms of MTE as follows:

$$\begin{aligned}
&\Delta^{TT}(x, D=1)\\
&=\frac{1}{\Pr(D=1|X=x)}\int\left[\int_{-\infty}^{z\beta_D}\mathrm{E}(\Delta|X=x, U_D=u)\,\mathrm{d}F_U(u)\right]\mathrm{d}F_{Z|X}(z|x)\\
&=\frac{1}{\Pr(D=1|X=x)}\int\left[\int\mathrm{E}(\Delta|X=x, U_D=u)\mathbf{1}[(u\leqslant z\beta_D)]\,\mathrm{d}F_{Z|X}(z|x)\right]\mathrm{d}F_U(u)\\
&=\int\mathrm{E}(\Delta|X=x, U_D=u)\frac{\Pr(D=1|X=x, U=u)}{\Pr(D=1|X=x)}\,\mathrm{d}F_U(u). \qquad (8)
\end{aligned}$$

Since $\Pr(D=1|X=x, U=u)/\Pr(D=1|X=x)$ is a non-increasing function of u, the TT parameter can be interpreted as a weighted average of marginal treatment effects where individuals who have unobserved characteristics that make them the most inclined to participate in the program (have low U_D values) receive the most weight in the average.

Heckman (1997), Heckman and Smith (1998) and Heckman et al. (1999) discuss the economic questions that these three parameters answer. In brief, MTE identifies the effect of an intervention on those induced to change treatment states by the intervention. TT estimates the effect of the program on the entire group of people who participate in it. ATE estimates the effect of the program on randomly selected persons. See Heckman and Vytlacil (2000) for a discussion of the relationships among these parameters.

Consider the special case where the outcome variable is dichotomous and is generated by an underlying linear latent index, $\mu_j(X, U_j)=\mathbf{1}[(X\beta_j\geqslant U_j)]$. In this special case, the mean treatment parameters have the following form:

$$\begin{aligned}
\Delta^{MTE}(x,u)&=\Pr(Y_1=1|X=x, U_D=u)-\Pr(Y_0=1|X=x, U_D=u)\\
&=F_{1|D}(x\beta_1|u)-F_{0|D}(x\beta_0|u),
\end{aligned}$$

$$\begin{aligned}
\Delta^{ATE}(x)&=\Pr(Y_1=1|X=x)-\Pr(Y_0=1|X=x)\\
&=F_{U_1}(x\beta_1)-F_{U_0}(x\beta_0),
\end{aligned}$$

$$\begin{aligned}
&\Delta^{TT}(x,z,D=1)\\
&=\Pr(Y_1=1|X=x, Z=z, D=1)-\Pr(Y_0=1|X=x, Z=z, D=1)\\
&=\frac{1}{F_{U_D}(z\beta_D)}[F_{D,1}(z\beta_D, x\beta_1)-F_{D,0}(z\beta_D, x\beta_0)],
\end{aligned}$$

$$\begin{aligned}
&\Delta^{TT}(x, D=1)\\
&=\Pr(Y_1=1|X=x, D=1)-\Pr(Y_0=1|X=x, D=1)\\
&=\frac{1}{\mathrm{E}(F_{U_D}(Z\beta_D)|X=x)}\,\mathrm{E}_Z[F_{D,1}(Z\beta_D, X\beta_1)-F_{D,0}(Z\beta_D, X\beta_0)|X=x],
\end{aligned}$$

where $F_{j|D}(t_j|t_D)=\Pr(U_j\leqslant t_j|U_D=t_D)$ for $j=0,1$. We now use the latent variable model to define distributional treatment parameters.

3.2. Distributional treatment parameters

For many questions, knowledge of distributional parameters is required.[9] Does anybody benefit from the program? Among those treated, what fraction is helped by the program and what fraction is hurt by it? We now consider treatment parameters for the distribution of treatment effects. We first define the distributional treatment parameters more generally, and then specialize to the case where the outcome variables are dichotomous and are generated by a latent index model. While there is a previous literature on the distribution of treatment effects,[10] this is the first analysis to define and analyze distributional treatment parameters in a manner parallel to the Heckman and Vytlacil (1999) analysis of mean treatment parameters.

For any measurable set $\mathscr{A}$, let $\mathbf{1}_{\mathscr{A}}(\zeta)$ be an indicator variable for the event $\zeta \in \mathscr{A}$. The parameters in Section 3.1 are defined as averages of Δ, and we can define the parallel parameters as averages of $\mathbf{1}_{\mathscr{A}}(\Delta)$ by simply substituting $\mathbf{1}_{\mathscr{A}}(\Delta)$ for Δ. Let $\mathscr{A}(x) = \{(u_0, u_1) : \mu_1(x, u_1) - \mu_0(x, u_0) \in \mathscr{A}\}$. Let $F_{0,1} = F_{U_0,U_1}$.

The distributional parameter corresponding to the *MTE* parameter for the event $\Delta \in \mathscr{A}$ is

$$\begin{aligned}\mathrm{E}[\mathbf{1}_{\mathscr{A}}(\Delta)|X = x, U_D = u, D = 1] &= \int \mathbf{1}_{\mathscr{A}}(\mu_1(x, u_1) - \mu_0(x, u_0))\, \mathrm{d}F_{0,1|D}(u_0, u_1|u) \\ &= \int_{\mathscr{A}(x)} \mathrm{d}F_{0,1|D}(u_0, u_1|u).\end{aligned}$$

A distributional parameter corresponding to the *ATE* parameter for the event $\Delta \in \mathscr{A}$ is

$$\begin{aligned}\mathrm{E}[\mathbf{1}_{\mathscr{A}}(\Delta)|X = x] &= \int \mathbf{1}_{\mathscr{A}}(\mu_1(x, u_1) - \mu_0(x, u_0))\, \mathrm{d}F_{0,1}(u_0, u_1) \\ &= \int_{\mathscr{A}(x)} \mathrm{d}F_{0,1}(u_0, u_1).\end{aligned}$$

Likewise, we can define a distributional parameter corresponding to the *TT* parameter conditional on Z for the event $\Delta \in \mathscr{A}$,

$$\begin{aligned}&\mathrm{E}[\mathbf{1}_{\mathscr{A}}(\Delta)|X = x, Z = z, D = 1] \\ &\quad = \frac{1}{\Pr(D = 1|Z = z)} \int \mathbf{1}_{(-\infty, z\beta_D]}(u_D)\mathbf{1}_{\mathscr{A}}(\mu_1(x, u_1) - \mu_0(x, u_0))\, \mathrm{d}F_{D,0,1}(u_D, u_0, u_1) \\ &\quad = \frac{1}{\Pr(D = 1|Z = z)} \int_{-\infty}^{z\beta_D} \left[\int_{\mathscr{A}(x)} \mathrm{d}F_{0,1|D}(u_0, u_1|u_D) \right] \mathrm{d}F_{U_D}(u_D),\end{aligned}$$

[9] Heckman (1992), Heckman et al. (1997b), and Heckman and Smith (1998) and Heckman and Vytlacil (2005) emphasize that many criteria for the evaluation of social programs require information on the distribution of treatment effects. See also Carneiro et al., (2001).

[10] See, e.g., Heckman (1992), Heckman et al. (1997b), Heckman and Smith (1998).

and a distributional parameter corresponding to the TT not conditioning on Z for the event $\Delta \in \mathscr{A}$,

$$\begin{aligned}
&\mathrm{E}[\mathbf{1}_{\mathscr{A}}(\Delta)|X=x, D=1] \\
&\quad = \frac{1}{\Pr(D=1|X=x)} \int \left(\int \mathbf{1}_{(-\infty, z\beta_D]}(u_D) \mathbf{1}_{\mathscr{A}}(\mu_1(x,u_1) - \mu_0(x,u_0)) \right. \\
&\qquad\qquad \left. \times \mathrm{d}F_{D,0,1}(u_D, u_0, u_1) \right) \mathrm{d}F_{Z|X}(z|x) \\
&\quad = \frac{1}{\Pr(D=1|X=x)} \int \left(\int_{-\infty}^{z\beta_D} \left[\int_{\mathscr{A}(x)} \mathrm{d}F_{0,1|D}(u_0, u_1|u_D) \right] \mathrm{d}F_{U_D}(u_D) \right) \mathrm{d}F_{Z|X}(z|x).
\end{aligned}$$

In the special case where the outcome variable is dichotomous and is generated by an underlying linear latent index, with $\mu_j(X, U_j) = \mathbf{1}[(X\beta_j \geqslant U_j)]$, Y_1 and Y_0 are binary and Δ can take three values. They are

(1) $\Delta = 1$ if the individual would have a successful outcome if treated (e.g., be employed if trained) and an unsuccessful outcome otherwise. ($Y_0 = 0$, $Y_1 = 1$).
(2) $\Delta = 0$ if the individual would have a successful outcome in either state ($Y_0 = 1$, $Y_1 = 1$), or the individual would have an unsuccessful outcome in either state ($Y_0 = 0$, $Y_1 = 0$).
(3) $\Delta = -1$ if the individual would have a successful outcome if not treated and an unsuccessful outcome if treated. ($Y_0 = 1$, $Y_1 = 0$).

Consider, for example, $\mathscr{A} = \{1\}$, so that $\mathbf{1}_{\mathscr{A}}(\Delta) = \mathbf{1}(Y_0 = 0, Y_1 = 1)$. In this case, $\mathscr{A}(x) = \{(u_0, u_1) : u_0 > x\beta_0, u_1 \leqslant x\beta_1\}$. We have

$$\begin{aligned}
\mathrm{E}[1_{\{1\}}(\Delta)|X=x] &= \Pr[Y_1 = 1, Y_0 = 0|X=x] \\
&= \Pr[Y_1 = 1|X=x] - \Pr[Y_1 = 1, Y_0 = 1|X=x] \\
&= F_1(x\beta_1) - F_{0,1}(x\beta_0, x\beta_1),
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}[1_{\{1\}}(\Delta)|X=x, Z=z, D=1] \\
&\quad = \Pr[Y_1 = 1, Y_0 = 0|X=x, Z=z, D=1] \\
&\quad = \frac{1}{\Pr[D=1|Z=z]} \Pr[Y_1 = 1, Y_0 = 0, D=1|X=x, Z=z] \\
&\quad = \frac{1}{\Pr[D=1|Z=z]} [\Pr[Y_1 = 1, D=1|X=x, Z=z] \\
&\qquad - \Pr[Y_1 = 1, Y_0 = 1, D=1|X=x, Z=z]] \\
&\quad = \frac{1}{F_{U_D}(z\beta_D)} [F_{D,1}(z\beta_D, x\beta_1) - F_{D,0,1}(z\beta_D, x\beta_0, x\beta_1)],
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}[1_{\{1\}}(\Delta)|X=x, D=1] \\
&\quad = \mathrm{E}_Z[\Pr[Y_1 = 1, Y_0 = 0|X=x, Z=z, D=1]|X=x, D=1]
\end{aligned}$$

$$= \frac{1}{\mathrm{E}_Z(F_{U_D}(Z\beta_D)|X=x)}\mathrm{E}_Z[F_{D,1}(Z\beta_D, X\beta_1) - F_{D,0,1}(Z\beta_D, X\beta_0, X\beta_1)|X=x],$$

$$\begin{aligned}&\mathrm{E}[1_{\{1\}}(\Delta)|X=x, U_D=u]\\ &\quad=\Pr[Y_1=1, Y_0=0|X=x, U_D=u]\\ &\quad=\Pr[Y_1=1|X=x, U_D=u]-\Pr[Y_1=1, Y_0=1|X=x, U_D=u]\\ &\quad=F_{1|D}(x\beta_1|z\beta_D)-F_{0,1|D}(x\beta_0, x\beta_1|z\beta_D)\quad \text{for } u=Z\beta_D.\end{aligned}$$

The corresponding parameters for $1_{\{-1\}}(\Delta)$ are defined by straightforward modification of the previous expressions. Notice that

$$\mathrm{E}(Y_1 - Y_0|X=x) = \mathrm{E}[1_{\{1\}}(\Delta)|X=x] - \mathrm{E}[1_{\{-1\}}(\Delta)|X=x]$$

so that the average treatment effect is the difference between two corresponding distributional parameters. The average gain (ATE) when outcome variables are binary is the probability of being successful (employed) when participating in the program minus the probability of being unsuccessful when participating in the program. Likewise, the other average treatment parameters can be seen as the difference between their corresponding distributional parameters for $1_{\{1\}}(\Delta)$ and $1_{\{-1\}}(\Delta)$. The distributional parameters offer a finer level of detail on the effectiveness of the program.

Identification of the distributional treatment parameters is more difficult than identification of the mean treatment parameters. Thus, identification of the bivariate distribution of (D, Y_1) and (D, Y_0) implies identification of the mean treatment parameters, while identification of the distributional treatment parameters requires knowledge of the full trivariate distribution of (D, Y_0, Y_1). Since Y_0 and Y_1 are never jointly observed, this trivariate distribution is not identified non-parametrically even when treatment is exogenous.[11]

However, the distribution of treatment effects can be identified if additional assumptions are made. We now discuss one such identifying assumption—that of a factor model.[12] A more systematic analysis of this assumption appears in Aakvik et al. (1999, Theorem 2) and Carneiro, Hansen and Heckman (2003).[13]

[11] See the discussion in Heckman (1990), Heckman et al. (1997b) and Heckman and Smith (1998).

[12] An alternative set of conditions sufficient to identify the full joint distribution (D, Y_0, Y_1) when Y_0, Y_1 are continuous involves using the Roy model with sufficient support conditions. See Heckman and Honoré (1990), Heckman (1990) and Heckman and Smith (1998). Aakvik et al. (1999) consider how the Roy structure can be used to identify the joint distribution in the context of dichotomous (Y_0, Y_1). Without imposing the factor structure assumption or the Roy model assumption, one can still bound the distribution of treatment parameters. This strategy is also explored in Aakvik et al. (1999). Carneiro et al., (2003) extend this analysis to a multiple choices, panel data setting with discrete and continuous outcome variables.

[13] If the factor structure is imposed, assumption (i)—the existence of an exclusion restriction on observables—need not be invoked. Some type of exclusion is required for a fully semiparametric analysis.

4. Factor structure models

In our empirical analysis we estimate a three equation model consisting of an equation for the decision rule, an outcome equation for the treated state, and an outcome equation for the non-treated state. The selection outcome and the employment outcomes are discrete. In this paper we specify a discrete-choice, latent-index framework where the unobservables are generated by a normal factor structure. Aakvik et al. (1999) consider more general, semiparametric cases.[14] The empirical results produced from the more general framework are in accord with the results reported here.

As before, the decision rule for training is[15]

$$D_i^* = Z_i\gamma - U_{Di},$$
$$D_i = 1 \text{ if } D_i^* \geqslant 0, \quad D_i = 0 \text{ otherwise.} \tag{9}$$

We specify the following employment outcome equation for the training state:

$$Y_{1i}^* = X_i\beta_1 - U_{1i},$$
$$Y_{1i} = 1 \text{ if } Y_{1i}^* \geqslant 0, \quad Y_{1i} = 0 \text{ otherwise,} \tag{10}$$

and the following employment outcome in the non-treated state:

$$Y_{0i}^* = X_i\beta_0 - U_{0i},$$
$$Y_{0i} = 1 \text{ if } Y_{0i}^* \geqslant 0, \quad Y_{0i} = 0 \text{ otherwise.} \tag{11}$$

We assume that the error terms in Eqs. (9)–(11) are governed by the following factor structure:

$$U_{Di} = -\theta_i + \varepsilon_{Di},$$
$$U_{1i} = -\alpha_1\theta_i + \varepsilon_{1i},$$
$$U_{0i} = -\alpha_0\theta_i + \varepsilon_{0i}. \tag{12}$$

The factor structure assumption for discrete choice models was introduced in Heckman (1981). It produces a flexible yet parsimonious specification which yields convenient and easily interpretable expressions for the parameters of interest and at the same time enables us to estimate the model in a tractable fashion.

We assume access to i.i.d. data, and henceforth suppress the i subscript. We make the following normality assumption,

$$\begin{pmatrix} \theta \\ \varepsilon_D \\ \varepsilon_1 \\ \varepsilon_0 \end{pmatrix} \sim N(0, I),$$

[14] Carneiro et al., (2003) extend this analysis to consider the case of multiple factor models.

[15] As is standard in economics, we impose linear index assumptions to reduce the dimensionality of the estimation problem. These parametric assumptions are not critical to our approach, and can be relaxed given a sufficiently large number of observations to make non-parametric estimation feasible. See Aakvik et al. (1999). Thus in our context, the application of Matzkin's framework (1992, 1994) is appealing.

where I is the identity matrix and where we have imposed the normalization that $\mathrm{Var}(\theta) = 1, \mathrm{Var}(\varepsilon_j) = 1$ for $j = D, 0, 1$.[16] This normalization is innocuous in the context of a normal factor model, see Heckman (1981) or Aakvik et al. (1999). Let Φ denote the standard normal CDF and let ϕ denote the standard normal probability density function.

The following expressions for the mean treatment parameters in the case of a normal factor model are easily verified:

$$\Delta^{ATE}(x) = \int [\Phi(x\beta_1 + \alpha_1\theta) - \Phi(x\beta_0 + \alpha_0\theta)]\phi(\theta)\,\mathrm{d}\theta,$$

$$\Delta^{TT}(x, z, D = 1) = \frac{1}{\Phi(z\beta_D/\sqrt{2})} \int [\Phi(x\beta_1 + \alpha_1\theta) - \Phi(x\beta_0 + \alpha_0\theta)]\Phi(z\beta_D + \theta)\phi(\theta)\,\mathrm{d}\theta,$$

$$\Delta^{TT}(x, D = 1) = \frac{1}{\mathrm{E}(\Phi(Z\beta_D/\sqrt{2})|X = x)} \mathrm{E}_Z \left(\int [\Phi(x\beta_1 + \alpha_1\theta) - \Phi(x\beta_0 + \alpha_0\theta)] \times \Phi(Z\beta_D + \theta)\phi(\theta)\,\mathrm{d}\theta | X = x \right),$$

$$\Delta^{MTE}(x, u) = \frac{\int (\Phi(x\beta_1 + \alpha_1\theta) - \Phi(x\beta_0 + \alpha_0\theta))\phi(u + \theta)\phi(\theta)\,\mathrm{d}\theta}{\phi(u/\sqrt{2})}.$$

Observe that if $\alpha_1 = \alpha_0$, we obtain a common treatment effect (conditional on X) for the indices of (10) and (11). However, we do not obtain a common treatment effect for the probability of employment gain.[17] Thus note that

$$\Delta^{ATE}(x) - \Delta^{TT}(x, z, D = 1) = \int [\Phi(x\beta_1 + \alpha_1\theta) - \Phi(x\beta_0 + \alpha_0\theta)] \left(1 - \frac{\Phi(z\beta_D + \theta)}{\int \Phi(z\beta_D + \theta)\phi(\theta)\,\mathrm{d}\theta} \right) \phi(\theta)\,\mathrm{d}\theta,$$

which will not in general equal zero unless $\alpha_1 = \alpha_0 = 0$.

The expressions for the distributional treatment parameters are easily derived. For example, the distributional parameters in this case for the event $\mathbf{1}[(Y_0 = 0, Y_1 = 1)] = \mathbf{1}[(\Delta = 1)]$ are

$$\mathrm{E}\mathbf{1}[(\Delta = 1)]|X = x] = \int [\Phi(x\beta_1 + \alpha_1\theta)(1 - \Phi(x\beta_0 + \alpha_0\theta))]\phi(\theta)\,\mathrm{d}\theta,$$

$$\mathrm{E}\mathbf{1}[(\Delta = 1)]|X = x, Z = z, D = 1] = \frac{1}{\Phi(z\beta_D/\sqrt{2})} \int [\Phi(z\beta_D + \theta)\Phi(x\beta_1 + \alpha_1\theta)(1 - \Phi(x\beta_0 + \alpha_0\theta))]\phi(\theta)\,\mathrm{d}\theta,$$

[16] An alternative approach assumes a discrete distribution of θ. In a finite sample, the NPMLE used by Heckman and Singer (1984), is a discrete distribution. See Cameron and Heckman (1987) where models with a discrete factor structure were first developed in the context of a discrete choice model.

[17] A common treatment effect conditional on X is a treatment effect which is the same for all persons with same X value.

$$\mathrm{E}\mathbf{1}[(\Delta = 1)]|X = x, D = 1]$$
$$= \frac{1}{\mathrm{E}(\Phi(Z\beta_D/\sqrt{2})|X = x)}\mathrm{E}\left[\int [\Phi(Z\beta_D + \theta)\Phi(x\beta_1 + \alpha_1\theta)[1 - \Phi(x\beta_0 + \alpha_0\theta)] \times \phi(\theta)\,\mathrm{d}\theta|X = x\right],$$

$$\mathrm{E}\mathbf{1}[(\Delta = 1)]|X = x, U_D = u] = \frac{\int \Phi(x\beta_1 + \alpha_1\theta)(1 - \Phi(x\beta_0 + \alpha_0\theta))\phi(u + \theta)\phi(\theta)\,\mathrm{d}\theta}{\phi(u/\sqrt{2})},$$

$$\mathrm{E}\mathbf{1}[(\Delta = 1)]|X = x, D = 1] = \frac{1}{\mathrm{E}(\Phi(Z\beta_D/\sqrt{2})|X = x)}\mathrm{E}\left[\int [\Phi(Z\beta_D + \theta) \times \Phi(x\beta_1 + \alpha_1\theta)[1 - \Phi(x\beta_0 + \alpha_0\theta)]\phi(\theta)\,\mathrm{d}\theta|X = x\right].$$

Observe that the random effects factor model of this section and the matching model of Rosenbaum and Rubin (1983) have a close affinity. If the analyst knew θ, then the matching conditions of Rosenbaum and Rubin (1983) would be satisfied.[18]

$$(Y_0, Y_1) \coprod D||X, Z, \theta,$$

where $\coprod$ denotes independence given the arguments to the right of the double bar, and

$$0 < \Pr(D = 1|X, Z, \theta) < 1,$$

where the latter assumption follows from the assumption that $\mathrm{Var}(\varepsilon_D) = 1$ and normality.[19] Thus given θ, we could use simple propensity score matching or other standard matching methods to estimate TT and ATE. However, matching does not identify MTE or the distributional parameters.[20]

Given that we do not observe θ, this strategy is not available to us or to many other analysts. Accordingly, we integrate out θ assuming that

$$\theta \coprod (X, Z).$$

[18] In matching, no exclusion restriction is required. There is no distinction between X and Z. Heckman et al. (1997a) introduce this distinction into matching.

[19] Under those conditions ε_D has full support on the real line.

[20] Heckman and Vytlacil (2005) develop this point at greater length. Neither the latent index model proposed here nor the matching method is more general than the other. For example, the method proposed here requires a first stage decision rule given by a threshold crossing model while the matching approach does not require this assumption. In contrast, the matching approach cannot allow for unobserved heterogeneity, while the method proposed here does allow for unobserved heterogeneity. Note that both latent index models and matching methods can be implemented either parametrically or non-parametrically, with the feasibility of non-parametric methods dictated by the amount of data available. See the exposition in Heckman and Navarro (2004).

Thus our random effects set up can be viewed as a solution to a missing conditioning variables problem in matching.[21]

Another approach to the problem of missing conditioning variable is to assume different values of the missing θ value and to perform a sensitivity analysis. This approach is advocated by Rosenbaum (1995, Chapter 5) and implemented in the context of a VR program by Aakvik (1999). We report estimates obtained from this procedure in Section 6 below.

5. Estimating the mixture model

Conditioning on θ, and restoring the i subscripts, the likelihood for the factor model has the form:

$$\prod_{i=1}^{N} \Pr(D_i, Y_i|X_i, Z_i, \theta_i),$$

where

$$\Pr(D_i, Y_i|X_i, Z_i, \theta_i) = \Pr(D_i|Z_i, \theta_i)\Pr(Y_i|D_i, X_i, \theta_i),$$

and

$$\Pr(D_i = 1|Z_i, \theta_i) = \Phi(Z_i\beta_D + \theta_i),$$

$$\begin{aligned}\Pr(Y_i = 1|D_i = 1, X_i, \theta_i) &= \Pr(Y_{1i} = 1|D_i = 1, X_i, \theta_i)\\ &= \Pr(Y_{1i} = 1|X_i, \theta_i)\\ &= \Phi(X_i\beta_1 + \alpha_1\theta_i),\end{aligned}$$

$$\begin{aligned}\Pr(Y_i = 1|D_i = 0, X_i, \theta_i) &= \Pr(Y_{0i} = 1|D_i = 0, X_i, \theta_i)\\ &= \Pr(Y_{0i} = 1|X_i, \theta_i)\\ &= \Phi(X_i\beta_0 + \alpha_0\theta_i).\end{aligned}$$

The likelihood function integrating out θ has the form

$$L = \prod_{i=1}^{N} \int \Pr(D_i, Y_i|X_i, Z_i, \theta)\phi(\theta)\,\mathrm{d}\theta.$$

Identification of the parameters of the model, $(\beta_D, \beta_0, \beta_1)$ and (α_0, α_1), follows from the analyses of Heckman (1981) or Aakvik et al. (1999) if $\varepsilon_D, \varepsilon_0, \varepsilon_1$ and θ are joint normal. We estimate the parameters by maximum likelihood, where we use Gaussian quadrature to approximate the integrated likelihood.[22]

[21] The random effects estimator is a member of the class of control function estimators discussed in Heckman and Vytlacil (2005).

[22] See Butler and Moffit (1982) for a discussion of Gaussian quadrature in this context. We use five evaluation points for the approximation, and we implement the maximum likelihood estimation using the DCPA package of Cameron and Heckman (1987).

The empirical results reported below are not sensitive to the assumption that θ is normally distributed. In alternative empirical analyses, we follow Heckman and Singer (1984) and Cameron and Heckman (1987) by approximating the distribution of θ with a distribution defined on a finite number of support points.[23] The empirical results obtained from using the discrete mixture model for θ are similar to the results generated by a normality assumption and for the sake of brevity are not reported.

Given identification of the parameters of the model, all mean and distributional treatment effect parameters are identified and standard errors for the treatment parameters follow from the delta method.[24] We integrate these estimated treatment parameters against the empirical distribution of X and Z to estimate the corresponding treatment parameters integrated over the distribution of X and Z. For example, we estimate $\mathrm{E}(\Delta)$ by $(1/N)\sum_{i=1}^{N}[F_{U_1}(X_i\beta_1) - F_{U_0}(X_i\beta_0)]$, where N is the sample size.

The assumption of a one-factor structure is crucial to the identification of distributional treatment effect parameters. The one factor structure implies that

$$\begin{aligned}
&\mathrm{Cov}(U_D, U_0) = \alpha_0,\\
&\mathrm{Cov}(U_D, U_1) = \alpha_1,\\
&\mathrm{Cov}(U_0, U_1) = \alpha_0\alpha_1.
\end{aligned}$$

(Recall we have scaled the variances of $\varepsilon_D, \varepsilon_0, \varepsilon_1$ and θ all to be one so that the normalizing constants are known). Thus, identification of α_0 (from $\mathrm{Cov}(U_D, U_0)$) and identification of α_1 (from $\mathrm{Cov}(U_D, U_1)$) immediately imply identification of $\mathrm{Cov}(U_0, U_1) = \alpha_0\alpha_1$. Given joint normality, this implies that the joint distribution U_D, U_0, U_1 is known. No exclusion restrictions are required (i.e. assumption (i) can be relaxed) and neither a Roy model structure is required (e.g. $D = 1(Y_1 \geqslant Y_0)$ nor its extension for latent variable models reported in Aakvik et al. (1999)). Replacing the normality assumption with non-parametric models requires some form of exclusion restriction. (see Aakvik et al., 1999; Carneiro et al., 2003).

[23] See also Cameron and Taber (1994) for a Monte Carlo analysis of the Cameron–Heckman model.

[24] More correctly, the standard errors of our estimators for the treatment parameters conditional on covariates follow from the delta method, since in all cases these estimators are smooth non-linear functions of the discrete choice parameter estimates. For example, the standard errors of our estimator of $\Delta^{ATE}(x)$ follows by the delta method. In contrast, the standard errors of our estimators of the unconditional treatment parameters require accounting for the sampling variability in the empirical distribution of the covariates. We can view our estimator of the treatment parameters not conditional on covariates as a multi-step estimation procedure, where the last step involves averaging over the covariates the estimated treatment parameter conditional on covariates. To derive the asymptotic theory for our estimators of the treatment parameters not conditional on covariates, one can follow Newey and McFadden (1994) in viewing a multi-step estimator as a GMM estimator for the stacked moment conditions with the identity matrix as the weighting matrix. Heckman et al. (2003) follow this strategy in a related, linear equation context.

6. Data and institutional setting

The Norwegian vocational rehabilitation sector offers income maintenance payments and training programs for individuals whose medical conditions result in reduced productivity. The VR sector has expanded rapidly since the National Social Insurance Act was passed in 1966. The expansion has been guided neither by a firm knowledge on the overall economic impact of the training programs, nor by knowledge of which groups may benefit most from program participation.[25] Today, around 1.5 percent of the labor force participates in a VR training program each day. Most persons who apply for VR job training programs have previously been employed.

Individuals unable to return to work after 52 weeks on sickness benefits are entitled to a VR benefit. The decision to provide the VR benefit is made by the local Social Security Office, usually after a recommendation from a medical doctor. The VR benefit is usually two-thirds of the gross income in the previous year subject to maximum and minimum benefit restrictions. Health status is the legal eligibility criterion for VR benefit.

While receiving a VR benefit, some people return to their old job or obtain disability pensions without entering training. Individuals who are not granted a disability pension and do not return to their old job on their own effort are usually referred to the local Employment Office to participate in a job training program. The office evaluates whether training may help applicants obtain a job.

The decision to accept a person into a training program is mainly taken by case workers at the Employment Office and by local managers of vocational rehabilitation centers. This decision is usually based on subjective judgments regarding employment prospects. Main inclusion criteria are health, age, personal characteristics, social conditions, education, and labor force experience. However, the vague criteria for selection and the close connection between the local labor market authorities and local firms and businesses may encourage case workers to select participants based on their expectations of post-program employment outcomes rather than on their expectations of post-program impacts.

The training programs offered are typically education (classroom training), formal on-the-job training in manufacturing sector firms, and wage subsidies.[26] The training

[25] There have been evaluation studies of rehabilitation programs in Sweden, see Heshmati and Engström (2001) and Frolich et al. (2004). However, there are large differences between the programs in the two countries. There is a clear administrative distinction between medical rehabilitation (MR) and vocational rehabilitation (VR) in Norway. Even though health improvements may occur during vocational rehabilitation, the main purpose of VR training programs in Norway is to enhance employability given medical diagnosis, not to improve health impairments. Swedish rehabilitation programs include both MR and VR and is thus not directly comparable to the Norwegian case. The Norwegian VR programs are organized and monitored by the Local Employment Office, so the link to the labor market is more direct in Norway compared to Sweden.

[26] We implement a two-state model (treatment or no-treatment), collapsing the different types of treatment into a single category. Given sufficient data, one could treat education, on-the-job training, and wage subsidies as separate treatments and thus have a four-state model. Data limitations prevent us from following this strategy.

Table 1
Means and standard deviations of characteristics for participants and non-participants

	1244 participants[a]	680 non-participants
	Means	
Employment rate, 1993	0.379 (0.485)	0.333 (0.472)
Income, 1988[b]	61,874 (54,420)	54,707 (55,324)
Income, 1992	60,509 (68,002)	52,076 (66,599)
One or more children[c]	0.562 (0.496)	0.520 (0.500)
Child older than 3[d]	0.459 (0.499)	0.390 (0.490)
Age	34.2 (10.1)	33.7 (10.7)
Actual years of work experience	9.7 (6.2)	9.6 (5.8)
Education in years	10.5 (1.8)	10.2 (2.0)
Degree of rationing[e]	0.17 (0.12)	0.26 (0.16)

[a] Participants are those individuals who applied for a training program in 1989 and were registered in a training program for at least 5 days in the period from the beginning of 1989 until the end 1993. Non-participants are those individuals who applied for a training program in 1989 but never registered as a participant in a training program.

[b] Income measured in 100,000 Kroner (NOK) for 1988 and 1992 using 1988 Kroner. 1988 is the year before the application to VR training programs.

[c] This is a dummy variable for having one or more children.

[d] This is a dummy variable for having a child over the age of three.

[e] Percentage of applicants in local districts who do not participate in VR training programs.

varies in substance and duration across clients. We would also expect the training effects to be heterogeneous because training is offered by different institutions. All schooling and labor market training are provided without direct cost for the participants, and participants usually receive a VR benefit while undergoing training. Even though the mean is around 6 months, this is not typical, because the variation is high.

We have a sample of 1,924 individuals who applied for training in 1989, which is a 10 percent random sample of all female VR clients who applied in 1989. Of these applicants, 1,244 were accepted and participated in a training program for at least 5 days. The remaining applicants were either not accepted into the training program or were accepted but chose not to participate.[27] For arguments in favor of "internal" comparison groups, see Bell et al. (1995), Heckman et al. (1998a). In brief, non-participants are located in the same labor market as participants and failure to match within local labor markets has been shown to be an important source of evaluation bias by Heckman et al. (1998a).

Table 1 contains descriptive statistics of the variables used in our empirical analysis for program participants and non-participants. The mean income in the year before application to the program is higher for participants than for non-participants. The average age of participants is half a year higher than that of

[27] We do not separately observe the case worker's acceptance decision and the decision of the applicant to attend training if accepted.

non-participants. Furthermore, participants have better education, and they are more likely to hold a job in 1993.[28] Aakvik (1999, 2001) provides a more comprehensive definition of our data source.

For the observables determining the treatment decision, we use a set of individual background characteristics as well as some aggregate variables calculated based on more than 400 municipalities. These administrative areas vary in population size (the mean is around 10,000) and geographic area. Our background variables include age, educational level, presence of children, age of the youngest child, income in the year before application to the program, and work experience as of the year before application to the program.[29] There is a marked difference between the probability of getting day-care placement for children below and above the age of three, and we have constructed a variable to capture this effect.[30]

Our instrument (Z) is the degree of rationing. This is calculated as the percentage of applicants in local districts who do not participate in the program. We expect the degree of rationing to influence a person's probability of participating in a training program, but not to affect the employment outcome after training. Unlike training programs for ordinary unemployed people, for which the number of training slots is correlated with the local unemployment rate, the number of slots for VR training programs does not vary due to changes in the unemployment rate in local districts, and instead depends only on the capacity of the local educational sector. Entry into the program is generated by health factors, which are only weakly related to local unemployment rates. The correlation between the degree of rationing variable and local unemployment rate is 0.01.[31] The availability of training slots thus appears to be a valid exclusion restriction.

We take the treatment decision (D_i) to be whether the applicant receives training (is both accepted into training and receives training).[32] All of our estimated treatment parameters are defined for the population of applicants. Thus, for example, the average treatment effect is the average treatment effect for individuals chosen at random from the pool of applicants, not from the pool of all eligible

[28] Employment is defined in this paper as working at least 20 hours per week at the end of our observation period, which is 1993. We have experimented with several different definitions of employment outcomes. For instance, we have used the full-time employed and conditioned on minimum spells of 60 and 90 days in a job in our definition of employment. However, the empirical results are not sensitive to our definition of employment.

[29] In this paper, age and previous income enter linearly into the index for the outcome equation. In an alternative specification, we included quadratic terms in age and previous income for the outcome equations. Age squared and previous income squared were not significantly different from zero in that specification. Including these terms improved the model fit only marginally, and had only a minor effect on the factor loadings.

[30] Surprisingly, spousal income plays only a minor role in participation decisions and we delete it in the final specifications.

[31] See Aakvik (1999).

[32] Since the selection process is a joint decision of the case worker and the client, it would be appropriate to follow Poirier (1980) in specifying a multiple index model. Given appropriate exclusion restrictions, our analysis can be extended to allow for a multiple index model of the selection process. We leave this extension to future work.

individuals.[33] We use employment three years after application to training as our outcome measure (Y_i).[34] While we observe employment for each year for 3 years after application to training, we only use the third year of the data since the employment status is highly correlated across post-application years. Specifying a panel data random coefficient training model requires special modeling due to the time dependence in outcomes. The required model is a natural extension of the framework developed in this paper (see Carneiro et al., 2003).

Many evaluation studies use earnings rather than employment as their outcome measure after training.[35] We use employment rates rather than earnings as our outcome measure for several reasons. First, public expenditure on VR programs is a part of the active labor market policy in Norway. This policy is intended to place as many people as possible into regular unsubsidized jobs, since the relatively generous social security system in Norway is likely to fail if high unemployment persists. Second, Norway has a relatively compressed wage distribution and we therefore expect that any effects of the program on earnings would be driven by effects of the program on employment rates.[36]

7. Estimating and interpreting mean treatment effects and distributional treatment effects

We now report estimates of the mean treatment parameters derived from the factor model presented in Section 4. We first discuss the estimated coefficient values. We then discuss what they imply for the various treatment effects. We then present our analysis of alternative definitions of cream-skimming and our evidence on this issue. Finally, we compare our estimates with those obtained from matching and linear instrumental variables methods. The model is estimated by maximum likelihood. Using chi-square measures of goodness of fit, the model fits the data well.[37]

As noted in Section 4, under our normality and factor structure assumptions, no exclusion restrictions or continuous regressors are required to identify the

[33] Whether the average treatment effect parameter is relevant to policy decisions depends on the particular application and the particular definition of the population of interest. In our case of the Norwegian VR program and defining the population of interest to be the pool of applicants, the average treatment effect parameter is particularly policy relevant. Program changes, such as increasing capacity to be able to train all applicants, or of requiring administrators to accept applicants randomly, are feasible policy options which the average treatment effect parameter directly addresses.

[34] Only 10 percent of applicants are still receiving VR benefits at our evaluation time (3 years after application to training).

[35] Card and Sullivan (1988) is a notable exception among non-experimental evaluations of US programs for its focus on employment rates as the outcome measure of interest. However, many non-experimental evaluations of European programs focus on employment outcomes. See Heckman et al. (1999) for a relevant survey.

[36] LaLonde (1995) and Heckman et al. (1999) point out that most of the earnings gains reported in the literature in the US follow from higher employment rates rather than from increased wages.

[37] Evidence on fit is available from the authors on request.

Table 2
Selection equation

	Coeff.	*t*-value	Marg.[a]
Factor	1.000		0.2378
Age	−0.018	2.34	−0.0042
Income[b]	0.021	2.53	0.0049
Education (years)	0.050	2.53	0.0117
One or more children[c]	−0.297	1.98	−0.0706
Child older than 3[d]	0.452	3.01	0.1075
Actual years of work exp.	0.010	0.93	0.0023
Degree of rationing[e]	−0.379	11.11	−0.0900

[a] Marginal effects are defined as the analytical derivative averaged over the unconditional distribution of Z: $E_Z(\partial \Pr(D = 1 \mid Z = z)/\partial z_k)$.

[b] Income measured in 100,000 Kroner (NOK). They are measured for 1988, the year before the application to VR training programs.

[c] This is a dummy variable for having one or more children.

[d] This is a dummy variable for having a child over the age of three.

[e] Percentage of applicants in local districts who do not participate in VR training programs.

mean or distributional treatment effects. Nonetheless, as noted in Section 6, in our data, we have a plausible exclusion restriction (a variable in Z but not in X): the degree of rationing. This variable is an important determinant of program participation. If this variable is not used, the fit of the model to the data substantially deteriorates.

7.1. Estimated coefficients

Estimates of the parameters of the selection equation (D^*), the employment equation for nonparticipants (Y_0^*), and the employment equation for participants (Y_1^*) are reported in Tables 2 and 3, respectively. For each equation, we report the parameter values, the *t*-values for the parameter values, and the mean marginal effects.[38] In both tables, only the results for the model with unobserved heterogeneity are reported. The estimates for the models that do not control for heterogeneity are available on request from the authors.

Consider the parameters related to selection into training. The estimated parameters of the selection equation reported in Table 2, Column 1 offer insight into the presence of non-random selection into rehabilitation programs. Individuals participating in the program differ significantly from eligible non-participants with respect to observable characteristics. If a potential participant had favorable characteristics associated with higher employment in either the trained or the untrained state before VR began, such as being young, having no

[38] Let x_k and z_k denote the kth element of X and Z, respectively. The mean marginal effects are defined as the analytical derivatives averaged over the unconditional distribution of either X or Z: $E_Z(\partial \Pr(D = 1|Z = z)/\partial z_k)$, $E_X(\partial \Pr(Y_0 = 1|X = x)/\partial x_k)$, and $E_X(\partial \Pr(Y_1 = 1|X = x)/\partial x_k)$.

Table 3
Employment equations

	Non-participation outcome			Participation outcome		
	Coeff.	*t*-value	Marg.[a]	Coeff.	*t*-value	Marg.[b]
Factor	0.433	1.28	0.1372	−0.307	0.92	−0.1072
Age	−0.042	4.28	−0.0042	−0.005	0.90	−0.0017
Income[c]	0.033	2.69	0.0103	0.000	2.45	0.0066
Education (years)	0.094	2.95	0.0297	0.107	5.13	0.0372
One or more children[d]	−0.769	3.35	−0.2440	0.006	0.04	0.0019
Child older than 3[e]	1.180	4.67	0.3744	0.108	0.72	0.0378
Actual years of work experience	0.077	4.91	0.0023	0.050	5.36	0.0174

[a] Marginal effects are defined as the analytical derivative averaged over the unconditional distribution of X: $\mathrm{E}_X(\partial \Pr(Y_0 = 1 \mid X = x)/\partial x_k)$.

[b] Marginal effects are defined as the analytical derivative averaged over the unconditional distribution of X: $\mathrm{E}_X(\partial \Pr(Y_1 = 1 \mid X = x)/\partial x_k)$.

[c] Income measured in 100,000 Kroner (NOK). They are measured for 1988, the year before the application to VR training programs.

[d] This is a dummy variable for having one or more children.

[e] This is a dummy variable for having a child over the age of three.

children and being well-educated, then he or she has a greater probability of participating in a training program. Given the presence of children, persons with older children are more likely to participate in a training program. The coefficient on the degree of rationing in local districts is statistically significantly different from zero.

Next turn to the employment equations. We report the estimated employment regression coefficients in Table 3, where the β_0-vector is reported in column 1 of Table 3, and the β_1 vector is reported in column 4 of Table 3. For both employment equations, all the estimated coefficients have reasonable signs and most of them are statistically significant. Young individuals with high education, no children, high working experience, and high previous income have the best chances of being employed at the end of our observation period. Young children decrease the probability of employment.

We also estimate the same model with the restriction that all factor loadings (α_0, α_1) equal zero. These results are available upon request. Fixing the factor loadings to zero imposes the restriction that the error terms are independent across equations, and thus does not allow for selection on unobservables related to the employment equations. The resulting estimates of the slope coefficients are similar to those reported in the tables for the more general models with factor loadings estimated. However, as discussed below in Section 7.6, imposing that the error terms are independent across equations results in a dramatic change in the estimated treatment parameter values. Moreover, the fit of the model to the data is slightly worse when we impose that the error terms are independent across equations. In this sense, a model with unobservables on which agents select is more consistent with the data.

Both likelihood ratio tests and t statistics on the factor loadings evaluated at conventional levels *do not* reject the null hypothesis of no selection bias due to unobservables. However, the estimated factor loadings are large even if imprecisely estimated.

In this paper we proceed conditional on the estimated non zero values of α. We are reluctant to set a large α_j to zero using a pretest estimator. Even if the reader wishes to set the α_j to zero based on the reported test statistics, the analysis that follows can be taken as an illustration of how to estimate a variety of interesting mean and distributional parameters if selection on unobservables is an empirically important phenomenon.[39]

7.2. Estimated mean treatment parameters

We next compute the different mean treatment parameters conditional on the maximum likelihood values for all of the parameters. We find that

$$ATE \equiv \mathrm{E}(\Delta) = -0.014$$

but it is not precisely estimated. Also,

$$TT \equiv \mathrm{E}(\Delta | D = 1) = -0.11.$$

For the entire population, the program has a slight negative effect, but has a stronger negative effect for those who are selected into the program. This suggests that selection into the program is perverse on net gains, a point we develop below in Section 7.5. In comparison, as reported in Table 1, the raw difference in mean outcomes is 0.046 ($\mathrm{E}(Y_1|D = 1) - \mathrm{E}(Y_0|D = 0) = 0.046$). Thus controlling for selection appears to affect the point estimates but the point estimates are not precisely estimated.

In order to study the relationship between unobservable characteristics related to program participation and the treatment effect, we plot the estimated MTE parameter for different values of U_D (see Fig. 1). The MTE parameter is increasing in U_D. It is negative for U_D values below 0.1, while it is large and positive for large U_D values. Recall that higher values of U_D imply lower probabilities in the program. Thus, in terms of unobservables, those most likely to participate benefit the least from the program. This evidence is consistent with our estimates for ATE and for the effect of treatment on the treated. From the analysis of Heckman and Vytlacil (2000, 2001, 2005), and from Eq. (8), the effect of treatment on the treated is an integrated version of MTE with most of the weight being placed on MTE values with small U_D values who are more likely to participate in the program. At these values MTE is very negative. ATE weights MTE more uniformly and accordingly is larger.

[39] The imprecision of our parameteric estimates suggests that a much larger sample sizes may be required to estimate the model non-parametrically.

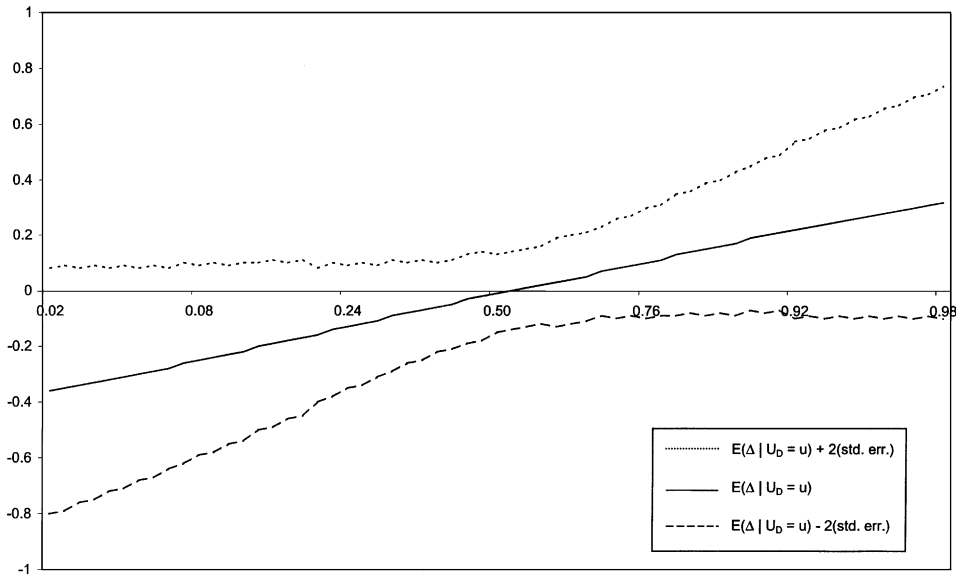

Fig. 1. Estimated marginal treatment effect.

7.3. Heterogeneity in observables

The estimated treatment effect vary substantially with observed characteristics. For example, the variance of $E(\Delta|X)$ is 0.0064 (standard error = 0.08), compared to its mean of −0.014. The variance of $E(\Delta|X, D = 1)$ is 0.0085 (standard error= 0.092) compared to its mean of −0.11. The degree to which the treatment effect varies with observable characteristics can also be seen by studying the marginal effect of each observable characteristic on the expected treatment effect. The marginal effects on the treatment parameters are reported in Table 4. For example, being older, having lower pre-program income, having lower spouse's income, and having young children are all associated with a larger treatment effect for all definitions of mean treatment effects. We develop this point further after we analyze distributional treatment parameters.

7.4. Estimated distributional treatment parameters

The distributional treatment effect parameters capture an additional type of treatment effect heterogeneity beyond that previously discussed for mean treatment effects. We now report estimates of the distributional treatment parameters. Table 5 reports the distributional versions of ATE, TT, and MTE evaluated at selected values of U_D. We find that if a random applicant is assigned to training, with probability 0.225 the applicant benefits from the training, that is, will be employed after receiving the training but would have been unemployed without the training. However, with probability 0.24 the applicant will be hurt by receiving the training,

Table 4
Marginal effects of regressors on mean treatment parameters

	$E_X\left[\frac{\partial E(\Delta \mid X=x)}{\partial x_K}\right]$	$E_X\left[\frac{\partial E(\Delta \mid X=x, D=1)}{\partial x_K} \mid D=1\right]$
Age	0.0115	0.0123
Income[a]	−0.0037	−0.0042
Education (years)	0.0075	0.0059
One or more children[b]	0.2459	0.2602
Child older than 3[c]	−0.3366	−0.3582
Actual years of work exp.[d]	−0.0070	−0.0083

[a] Income measured in 100,000 Kroner (NOK) and measured for the year before the application to VR training programs.
[b] This is a dummy variable for having one or more children.
[c] This is a dummy variable for having a child.
[d] Percentage of applicants in local districts who do not participate in VR training programs.

being unemployed 3 years after receiving the training but employed without receiving the training.[40] The mean parameter for ATE, −0.014, masks the underlying heterogeneity that nearly a quarter of all individuals become employed as a consequence of participating in the program and nearly a quarter of all individuals who participate become unemployed. The impact of the program is most negative for those most likely to enter the program. For example, the estimated MTE parameter evaluated at $U_D = -2$ shows that only 11.9% of such agents become employed because of participation but 37.3% of such agents become unemployed because of the program. These numbers are reversed for those least likely to enter the program (high values of U_D). The MTE parameter evaluated at $U_D = 2$ finds 35% of agents with $U_D = 2$ become employed because of participation and only 12.6% of such agents become unemployed due to participation in the program. As a consequence of our index model, we can write $\Pr(\Delta = 1 \mid X) = \Pr(\Delta = 1 \mid X\beta_1, X\beta_0)$ so that two indices capture the full X effect on $\Pr(\Delta = 1 \mid X)$. Fig. 2 graphs this function against quantiles of $X\beta_1$ and $X\beta_0$. Figs. 3 and 4 are the corresponding graphs for $\Pr(\Delta = 0 \mid X\beta_1, X\beta_0)$ and $\Pr(\Delta = -1 \mid X\beta_1, X\beta_0)$, respectively. The greatest gains in employment are for those with the highest index for employment in the treated state and the lowest employment index for the untreated state. The graph for $\Pr(\Delta = -1 \mid X\beta_1, X\beta_0)$ is the mirror image of Fig. 2. The graph for $\Pr(D = 0 \mid X\beta_1, X\beta_0)$ shows that persons with balanced levels of indices are the ones most likely to be unaffected by the program ($\Delta = 0$). There is considerable heterogeneity in response to the program among persons with different X values.

[40] One possible reason for the negative effect of the training on some individuals is a stigma effect associated with the training program. In interviews with the program administrators conducted by the authors, many of the administrators expressed concern that the receipt of the VR training conveys a negative signal to potential employers. For example, some of the administrators believed that receipt of VR training acts as a signal to potential employers that the trainee is prone to long spells of sickness.

Table 5
Mean and distributional treatment parameters

ATE $E(\Delta) = -0.014$ (standard error = 0.08)	Distributional version of *ATE*: $Pr[\Delta = 1] = 0.225$ $Pr[\Delta = 0] = 0.532$ $Pr[\Delta = -1] = 0.240$
TT $E(\Delta \mid D = 1) = -0.110$ (standard error = 0.09)	Distributional version of *TT*: $Pr[\Delta = 1 \mid D = 1] = 0.178$ $Pr[\Delta = 0 \mid D = 1] = 0.534$ $Pr[\Delta = -1 \mid D = 1] = 0.288$
MTE with $U_D = 2$ $E(\Delta \mid U_D = 2) = 0.224$ (standard error = 0.17)	Distributional version of *MTE* with $U_D = 2$: $Pr[\Delta = 1 \mid U_D = 2] = 0.350$ $Pr[\Delta = 0 \mid U_D = 2] = 0.524$ $Pr[\Delta = -1 \mid U_D = 2] = 0.126$
MTE with $U_D = 0$ $E(\Delta \mid U_D = 0) = -0.014$ (standard error = 0.07)	Distributional version of *MTE* with $U_D = 0$: $Pr[\Delta = 1 \mid U_D = 0] = 0.219$ $Pr[\Delta = 0 \mid U_D = 0] = 0.549$ $Pr[\Delta = -1 \mid U_D = 0] = 0.233$
MTE with $U_D = -2$ $E(\Delta \mid U_D = -2) = -0.255$ (standard error = 0.16)	Distributional version of *MTE* with $U_D = -2$: $Pr[\Delta = 1 \mid U_D = -2] = 0.119$ $Pr[\Delta = 0 \mid U_D = -2] = 0.508$ $Pr[\Delta = -1 \mid U_D = -2] = 0.373$

7.5. Cream-skimming: the relationship between selection into the program and outcomes

A central question in the analysis of a program like VR is whether those who benefit the most from it are those most likely to participate in it. We have already noted that *ATE* is greater than *TT*, i.e., that randomly selected persons benefit more from the program than those who participate in it. This suggests that the combinations of U_D and Z values that promote program participation are perversely associated with the observed and unobserved factors associated with gains from the program.

In order to determine the extent of cream-skimming on both observables and unobservables, it is necessary to relate Δ (as defined by the various means and distributional parameter analogues) to $Z\beta_D$ and U_D. We have estimated relationships among Δ and $(X\beta_1, X\beta_0, U_1, U_0)$, however. So the problem is how to go from the relationships we have estimated to determine the relationships between gains and $Z\beta_D$ and U_D.

Given the factor structure model, we can easily determine how variation in U_D affects U_1 and U_0 (see Eq. (12)). By virtue of independence assumption (iii), the factor relationship does not depend on values of $Z\beta_D$, $X\beta_1$ and $X\beta_0$. We have used this relationship in computing Fig. 1 and in inferring that selection into the program

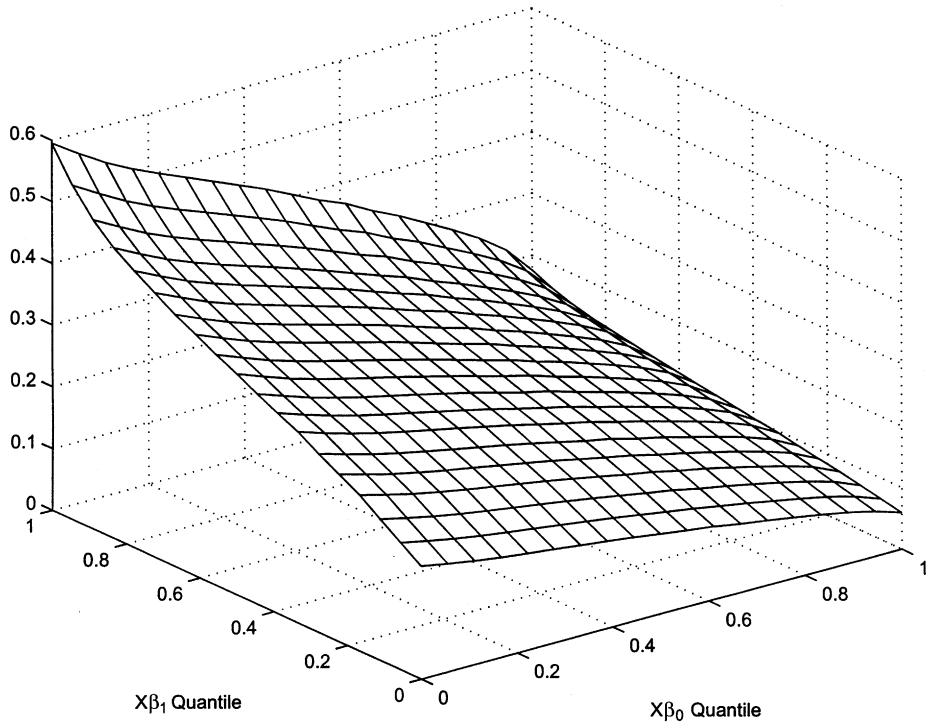

Fig. 2. $\Pr(\Delta = 1 \mid X\beta_1, X\beta_0)$.

is perverse in terms of unobservables. Another way to make the same point is to inspect the correlations among the unobservables. Using our normalizations

$$Corr(U_0, U_1) = \frac{\alpha_0\alpha_1}{\sqrt{1+\alpha_0^2}\sqrt{1+\alpha_1^2}} = -0.116,$$

$$Corr(U_D, U_0) = \frac{\alpha_0}{\sqrt{2}\sqrt{1+\alpha_0^2}} = 0.281,$$

$$Corr(U_D, U_1) = \frac{\alpha_1}{\sqrt{2}\sqrt{1+\alpha_1^2}} = -0.208.$$

From the first correlation, we have that the unobservables determining employment status in the no-training and training states are only weakly correlated. The point estimate is negative, suggesting that those individuals with unobservables which make them more likely to be employed in the training state are slightly less likely to be employed in the no-training state. From the latter two correlations, the unobservables that promote participation are positively correlated with the unobservables that promote employment in the no-training state but are negatively correlated with the unobservables that promote employment in the training state.

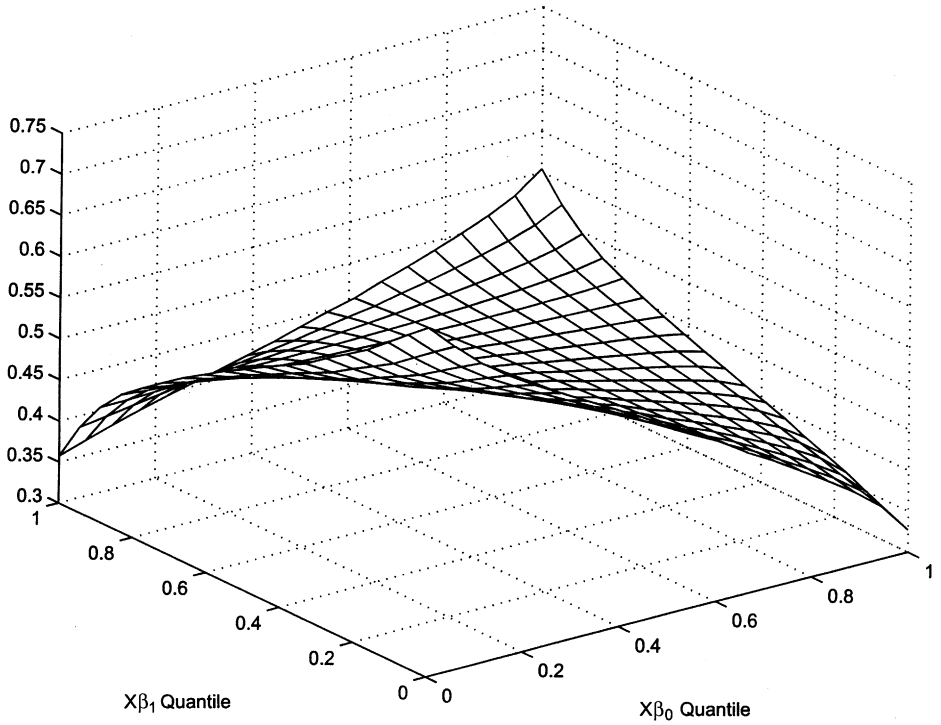

Fig. 3. $\Pr(\Delta = 0 \mid X\beta_1, X\beta_0)$.

Thus higher U_D is associated with higher U_0 and lower U_1 so that persons with low values of U_D (who are more likely to participate in the program) are more likely to have lower values of Δ, holding constant X and Z. Hence, selection is perverse on unobservables: treatment effects are the lowest for those most likely to participate.

The harder problem is to determine the effect of $Z\beta_D$ on Δ. The obvious way to assess this dependence is to estimate our model non-parametrically, determining the relationship between objects like ATE and Treatment on the Treated on $Z\beta_D$. A completely general way to express ATE in terms of $Z\beta_D$ writes

$$\mathrm{E}(\Delta \mid Z\beta_D) = \mathrm{E}_{Z\beta_D}[\mathrm{E}(\Delta \mid X)] = \int \mathrm{E}(Y_1 - Y_0 \mid X = x)\, \mathrm{d}F(x \mid Z\beta_D).$$

A comparable expression can be derived for $\mathrm{E}(\Delta \mid Z\beta_D, D = 1)$ the TT parameter:

$$\mathrm{E}(\Delta \mid Z\beta_D = z\beta_D, D = 1) = \int \mathrm{E}(Y_1 - Y_0 \mid X = x, Z\beta_D = z\beta_D, D = 1)\, \mathrm{d}F(x \mid D = 1, Z\beta_D = z\beta_D).$$

To estimate these expressions requires determining the distributions of $F(x \mid Z\beta_D = z\beta_D)$ and $F(x \mid Z\beta_D = z\beta_D, D = 1)$. The effect of the X can be reduced to the effect of two scalars, $(X\beta_1, X\beta_0)$, by virtue of our index assumption. To estimate these

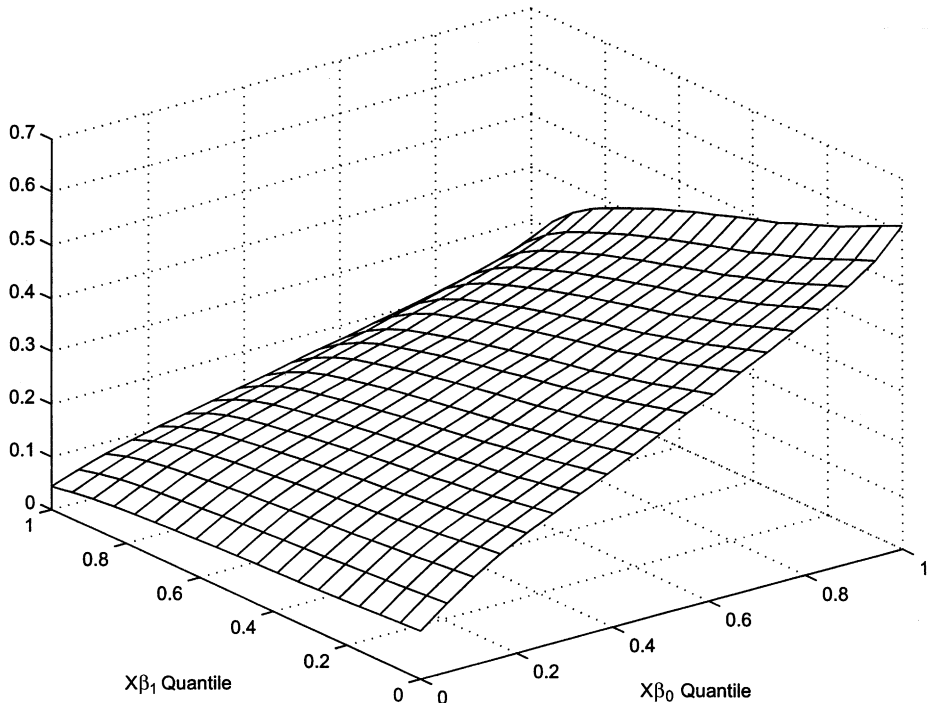

Fig. 4. $\Pr(\Delta = -1 \mid X\beta_1, X\beta_0)$.

densities requires non-parametric estimation of bivariate densities, a task we leave for another occasion.

Instead, we examine the dependence among the indices $(Z\beta_D, X\beta_0, X\beta_1)$ using correlations. We view this analysis as a prelude to a full non-parametric analysis. Many of the same characteristics that predict employment in the non-participation state also predict employment in the participation state. Also, many of the same characteristics that predict participation in training also predict higher employment probabilities for both the participation and non-participation outcome equations. We estimate the following correlations across the indices of these equations:

$$Corr(X\beta_0, X\beta_1) = 0.81,$$
$$Corr(Z\beta_D, X\beta_0) = 0.42,$$
$$Corr(Z\beta_D, X\beta_1) = 0.27.$$

The indices are all positively correlated with one another. Thus, unlike the case that arises in our analysis of unobservables, a higher index for participation is associated with higher employment outcomes in both the treated and untreated states and the effect on Δ depends on the levels at which the indices are related. Note that the correlation between $X\beta_1$ and $X\beta_0$ is strong and positive but they are not perfectly correlated. There is a strong relationship between observable characteristics that

predict participation and observable characteristics that predict employment in the non-participation state. $Corr(Z\beta_D, X\beta_0) = 0.43$. The correlations in the indices induce very similar correlations in the fitted probabilities:

$$Corr(\Pr(Y_1 = 1 \mid X), \Pr(Y_0 = 1 \mid X)) = 0.81,$$
$$Corr(\Pr(D = 1 \mid Z), \Pr(Y_0 = 1 \mid X)) = 0.42,$$
$$Corr(\Pr(D = 1 \mid Z), \Pr(Y_1 = 1 \mid X)) = 0.27.$$

The correlations combining both the observable and unobservable components of the indices are

$$Corr(Y_1^*, Y_0^*) = Corr(X\beta_1 + U_1, X\beta_0 + U_0) = 0.05,$$
$$Corr(D^*, Y_0^*) = Corr(Z\beta_D + U_D, X\beta_0 + U_0) = 0.31,$$
$$Corr(D^*, Y_1^*) = Corr(Z\beta_D + U_D, X\beta_1 + U_1) = -0.14.$$

Note that the correlation between the latent index for participation and the latent index for employment in the non-participation state is even higher than the correlation between the indices for employment in the participation and non-participation states.

The difference between the mean outcomes and the selection-corrected mean outcomes is consistent with the evidence just discussed. In particular, individuals who are most likely to enter the program are those who are most likely to be employed. In addition, those with the characteristics that make them most likely to participate are the ones who benefit the least from the program. This is true both for the observed characteristics and the unobserved characteristics. In terms of observed characteristics, note that

$$Corr(Z\beta_D, X(\beta_1 - \beta_0)) = -0.41$$

which induces a similar correlation between the fitted probabilities of participation and the expected treatment effect,

$$\begin{aligned} Corr(\Pr(D = 1 \mid Z), \mathrm{E}(\Delta \mid X)) &= Corr(\Pr(D = 1 \mid Z), \Pr(Y_1 = 1 \mid X) - \Pr(Y_0 = 1 \mid X)) \\ &= -0.38. \end{aligned}$$

In terms of unobserved characteristics,

$$Corr(U_D, U_1 - U_0) = -0.33.$$

The correlation in the observables and unobservables reinforce one another, resulting in

$$Corr(D^*, Y_1^* - Y_0^*) = -0.33.$$

This analysis demonstrates that those most likely to participate in the program are those who benefit the least from it. Contrary to several US studies which find that persons with characteristics associated with better labor market outcomes also gain the most from training (see the studies summarized in Heckman et al. (1999)), we find that characteristics associated with better labor market outcomes are negatively correlated with training effects.

Table 6
Effects of Norwegian VR training on employment (standard errors in parentheses)

Unconditional mean differences[a]	0.046	
	(0.023)	
	ATE	*TT*
Matching on observables[b]	0.043	0.028
	(0.023)	(0.019)
Model without unobserved Heterogeneity[c]	0.035	0.028
	(0.023)	(0.022)
Model with normally distributed unobserved Heterogeneity[d]	−0.014	−0.110
	(0.080)	(0.092)
Linear IV—common treatment effect[e]	−0.004	−0.004
	(0.078)	(0.078)
Linear IV—treatment[f] allowed to vary with X	0.015	0.012
	(0.043)	(0.031)

[a] Unconditional mean difference is $\hat{E}(Y_1 \mid D = 1) - \hat{E}(Y_0 \mid D = 0)$.
[b] Matching estimates taken from Aakvik (1999). These estimates are based on interval matching on the propensity score.
[c] Model without unobserved heterogeneity is based on latent variable model with $\alpha_0 = \alpha_1 = 0$.
[d] Model with normally distributed unobserved heterogeneity is based on latent variable model with normal factor structure.
[e] Linear IV—common treatment effect is based on a LPM form for the outcome equation and a common treatment effect assumption ($Y = X\beta + \gamma D + U$) and linear IV estimation using $\Pr(D = 1 \mid Z)$ as the instrument for D.
[f] Linear IV—treatment is based on a LPM form for the outcome equation while allowing the treatment effect to depend on observables ($Y = X\beta_0 + D(X\beta_1 - X\beta_0) + U$), with $\Pr(D = 1 \mid Z)X$ used as the instrument for DX.

The overall effectiveness of VR training programs in terms of producing employment gains can be improved by changing the selection criteria for participating in training. By focusing on selecting persons with a high training effect, the mean effect of VR training can be improved, although the gross employment outcomes among participants may be reduced.[41]

7.6. Comparison with results using other estimation methods

In Table 6, we compare our estimated treatment parameters with the estimates produced from alternative estimators. In particular, we compare our estimated *ATE* and *TT* parameters with the estimates resulting from: (1) mean difference in outcomes between participants and non-participants (i.e., assuming treatment is exogenous); (2) estimation by matching on observables, using the analysis of Aakvik (2001), (3) estimating our latent index model but with the factor loadings set to zero; (4) estimating our model using normal heterogeneity; (5) estimation by linear IV,

[41] We lack information on costs. Thus the net social benefit of our proposed change in the strategy of selection may be negative. See Heckman and Smith (1998) for evidence on the importance of accounting for full social costs in evaluating social programs.

where we impose a linear probability model form for the outcome equation assuming that treatment only shifts the intercepts of the outcome equations, and use the estimated $\Pr[D = 1 \mid Z]$ as the instrument for D; (6) estimation by linear IV, following a suggestion of Angrist (2001), where we impose a linear probability model form for the outcome equation, allow the treatment effect to vary with observable variables, and use the estimated $\Pr[D = 1 \mid Z] \times X$ as the instrument for DX.[42] Note that each of these estimators is based on different identifying assumptions. (see Heckman and Vytlacil (2005) for a discussion of these assumptions).

Notice the following features of Table 6. First, the raw difference in mean outcomes is higher than the estimated ATE or TT from any of the other estimators. Second, controlling for unobservables affects the estimates. The estimators that allow for selection on unobservables all produce estimates of TT that are somewhat lower than the estimates produced by estimators that assume no selection on unobservables like the IV estimators. However, for ATE the two types of estimators are in closer agreement. Third, for each estimator that allows ATE and TT to be distinct, it is always the case that TT is lower than ATE. Fourth, matching and a version of our model without unobserved heterogeneity produce essentially the same estimates, exactly as the theory developed in Section 4 predicts. We stress that given the imprecision of the estimates, few sharp distinctions can be maintained.

7.7. Placing our empirical results in the literature on VR programs

The literature on vocational rehabilitation programs reports no estimates of training effects based on randomized experiments. The literature on manpower programs directed towards the unemployed contains both methodological and empirical discussions of the relative merits of experimental and non-experimental evaluation methods.[43]

Methodologically, the problems of evaluating manpower and VR training programs are quite similar. However, it is not clear that the empirical regularities in the manpower literature apply to VR programs because participants in VR training programs have health problems that distinguish them from healthy individuals who are unemployed.

Very few studies of the effects of VR training control for potentially successful rehabilitation without training at all. In a review of the literature analyzing US data, Worrall (1988) focuses on this shortcoming. He notes that all the studies that he reviews are hampered by the lack of a control group. Nevertheless, other researchers have attempted to draw some inference from the same literature that Worrall reviews. Haveman et al. (1984) offer a guarded assessment, stating that concentrating rehabilitation activities on younger, less disabled, and more productive workers appears to be more efficient in promoting employment than focusing on disabled workers with the most severe handicaps.

[42] Moffit (2001) and Todd (2001) evaluate the merits of this ad hoc procedure.

[43] See, e.g., Heckman and Hotz (1989), Burtless (1995), Heckman and Smith (1995), LaLonde (1995), and Heckman et al. (1999).

Nowak (1983) and Dean and Dolan (1991) analyze the effects of VR programs in the US using a comparison group approach. Both these analyses find evidence that suggests a difference between gross success rates and training effects. Both studies report that females benefit more from training than males.[44] In results available on request from the authors, gender effects in Norwegian VR training programs are opposite to those found in the United States: estimated training effects are higher for males. We also find a statistically significant selection of young, relatively highly educated individuals with long work experience and no children into training programs. These individuals have a significantly higher probability of employment, with or without the treatment. Another characteristic that is predictive of a significant and positive effect on employment rates is high yearly pre-program income. Thus, as noted in Section 6, program managers appear to select participants so as to maximize gross employment rates among participants in training programs.

Our analysis suggests that a different selection rule might increase the overall efficacy of training in promoting gains in employment, assuming that the costs are the same across different selection rules. By concentrating on older, less educated women with low levels of work experience, there would be a drop in their recorded employment rates, but an increase in employment attributable to the program.

8. Summary, conclusions and related work

This paper formulates an econometric framework for studying treatment effects on discrete outcomes when the treatment effects vary among observationally identical persons. Using a latent variable framework, we show how to define and estimate the average treatment effect, the effect of treatment on the treated, and the marginal treatment effect on discrete outcomes. We also develop and estimate distributional analogues to these parameters. To secure estimates, we assume a factor-structure assumption for the model unobservables.

In related research, we relax the parametric normal assumptions used in this paper to construct semiparametric mean and distributional parameters. We present formal proofs of identification and a sampling theory for the semiparametric estimators considered in that paper. (see Aakvik et al., 1999). Carneiro et al., (2003) extend our framework by (a) analyzing panel data, (b) allowing for multiple factors, (c) adjoining measurement systems to outcome equations to identify factors and innovations non-parametrically, and (d) considering more general choice processes.

The Norwegian VR program we study offers different general and specific training programs at different locations to a diverse population. The estimated effects of these training programs vary both in terms of observed and unobserved factors. In

[44] Similar results are found in several other training program evaluations, see the reviews by Barnow (1987), Gueron and Pauly (1991), LaLonde (1995), and Heckman et al. (1999). Typically, they find that training has a significant effect for women, and that there are no significant effects for men and youths in terms of increased employment rates and wages. These results apply both for experimental and for observational studies.

particular, training effects appear to be larger for individuals with characteristics that predict lower employment in either the trained or untrained state. Cream-skimming of individuals into training on the basis of characteristics that are positively associated with employment is less effective in promoting employment gains than randomly selecting participants from the pool of applicants. There appears to be potential for improving the overall employment-promoting effect of VR training by selecting those who gain the most from training rather than choosing more employable persons.

Governmental evaluations of training programs in most countries typically are based on post-program outcome measures. Such an evaluation strategy gives caseworkers an incentive to select the most employable for training. Caseworkers are seldomly able to estimate treatment effects. Thus, guidance on who should participate should be based on results from research rather than by rules-of-thumb. We find that the employment gains in the Norwegian VR program will be enhanced if the selection rule is changed to encourage the least employable to participate.

Acknowledgements

We thank Victor Aguirregabiria, Xiaohong Chen, Jean-Pierre Florens, Lars Hansen, Robert LaLonde, and Costas Meghir for helpful comments.

References

Aakvik, A., 1999. Five essays on the microeconometric evaluation of job training programs. Dissertation, University of Bergen.

Aakvik, A., 2001. Bounding a matching estimator: the case of a Norwegian training program. Oxford Bulletin of Economics and Statistics 63 (1), 115–143.

Aakvik, A., Heckman, J., Vytlacil, E., 1999. Semiparametric program evaluation: lesson from an evaluation of a Norwegian training program. University of Chicago, unpublished manuscript.

Anderson, K.H., Burkhauser, R.V., Raymond, J.E., 1993. The effect of creaming on placement rates under the job training partnership act. Industrial and Labor Relations Review 46 (4), 613–624.

Angrist, J., 2001. Estimation of limited dependent variables with binary endogenous regressors: simple strategies for empirical practice. Journal of Economics and Business Statistics (see also discussions by Moffit and Todd) 19, 2–28.

Angrist, J., Graddy, K., Imbens, G., 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Review of Economic Studies 67, 499–527.

Barnow, B., 1987. The impact of CETA programs on earnings: a review of the literature. Journal of Human Resources 22, 157–193.

Bassi, L.J., 1983. The effect of CETA on the postprogram earnings of participants. Journal of Human Resources 18 (4), 539–556.

Bell, S., Orr, L., Blomquist, J., Cain, G., 1995. Program applicants as a comparison group in evaluating training programs. W. E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Björklund, A., Moffit, R., 1987. The estimation of wage gains and welfare gains in self-selection models. Review of Economics and Statistics 69, 42–49.

Burtless, G., 1995. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives 9 (2), 63–84.

Butler, J., Moffit, R., 1982. A computationally efficient quadrature procedure for the one factor multinomial probit model. Econometrica 50, 761–764.
Cameron, S., Heckman, J., 1987. Son of CTM: the DCPA approach based on discrete factor structure models. Working Paper, University of Chicago.
Cameron, S., Taber, C., 1994. Evaluation and identification of semiparametric maximum likelihood models of dynamic discrete choice. Working Paper, University of Chicago.
Card, D., Sullivan, D., 1988. Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 56, 497–530.
Carneiro, P., Hansen, K., Heckman, J., 2001. Removing the veil of ignorance in assessing the distributional impacts of social policies. Swedish Economic Policy Review 8, 273–301.
Carneiro, P., Hansen, K., Heckman, J., 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. International Economic Review 44, 361–422.
Cox, D.R., 1958. The Planning of Experiments. Wiley, New York.
Dean, D., Dolan, R., 1991. Fixed-effects estimates of earnings impacts for the vocational rehabilitation program. Journal of Human Resources 26, 380–391.
Fisher, R.A., 1935. Design of Experiments. Oliver and Boyd, London.
Frolich, M., Heshmati, A., Lechner, M., 2004. A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. Journal of Econometrics, Forthcoming.
Gay, R., Borus, M., 1980. Validating performance indicators for employment and training programs. Journal of Human Resources 15, 29–48.
Gritz, M., 1993. The impact of training on the frequency and duration of employment. Journal of Econometrics 57, 21–51.
Gueron, J., Pauly, E., 1991. From Welfare to Work. Russell Sage Foundation, New York.
Ham, J., Lalonde, R., 1996. The effects of sample selection and initial conditions in duration models: evidence from experimental data on training. Econometrica 64, 175–205.
Haveman, R., Halberstadt, V., Burkhauser, R., 1984. Public Policy Towards Disabled Workers: Cross-National Analyses of Economic Impacts. Cornell University Press, Ithaca, New York.
Heckman, J., 1981. Statistical models for discrete panel data. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.
Heckman, J., 1990. Varieties of selection bias. American Economic Review 80, 313–318.
Heckman, J., 1992. Randomization and social program evaluation. In: Manski, C., Garfinkle, I. (Eds.), Evaluating Welfare and Training Program. Harvard University Press, Cambridge, MA, pp. 201–230.
Heckman, J., 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. Journal of Human Resources 32, 441–462.
Heckman, J., Honoré, B., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149.
Heckman, J., Hotz, J., 1989. Choosing among alternative methods of estimating the impact of social programs: the case of manpower training. Journal of the American Statistical Association 84, 862–874.
Heckman, J., Navarro, S., 2004. Using matching, instrumental variables and control functions to estimate economic choice models. Review of Economics and Statistics 86 (1), 30–57.
Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York, pp. 156–245.
Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration models. Econometrica 52, 271–320.
Heckman, J., Smith, J., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9, 85–100.
Heckman, J., Smith, J., 1998. Evaluating the welfare state. In: Strom, S. (Ed.), Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial, Econometric Society Monograph Series. Cambridge University Press, Cambridge.
Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96, 4730–4734.

Heckman, J., Vytlacil, E., 2000. The relationship between treatment parameters within a latent variable framework. Economic Letters 66, 33–39.

Heckman, J., Vytlacil, E., 2001. Local instrumental variables. In: Hsiao, C., Morimune, K., Powell, J. (Eds.), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge, pp. 1–46.

Heckman, J., Vytlacil, E., 2005. Econometric evaluations of social programs. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, Vol. 5. North-Holland, Amsterdam, forthcoming.

Heckman, J., Ichimura, H., Todd, P., 1997a. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. Review of Economic Studies 64, 605–654.

Heckman, J., Smith, J., Clements, N., 1997b. Social experiments: accounting for heterogeneity in programme impacts. Review of Economic Studies 64, 487–535.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.

Heckman, J., Lochner, L., Taber, C., 1998b. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Review of Economic Dynamics 1 (1), 1–58.

Heckman, J., Lochner, L., Taber, C., 1998c. General equilibrium treatment effects: a study of tuition policy. American Economic Review 88 (2), 381–386.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of training programs. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. III. North-Holland, Amsterdam.

Heckman, J., Heinrich, C., Smith, J., 2002. The performance of performance standards. Journal of Human Resources 37 (4), 778–811.

Heckman, J., Tobias, J., Vytlacil, E., 2003. Simple estimators for treatment parameters in a latent variable framework. Review of Economics and Statistics 85 (3), 748–755.

Heshmati, A., Engstrom, L.G., 2001. Estimating the effects of vocational rehabilitation programs in Sweden. In: Michael, L., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica, Heidelberg, pp. 183–210.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–476.

Junker, B., Ellis, J., 1997. A characterization of monotone unidimensional latent variable models. Annals of Statistics 25, 1327–1343.

LaLonde, R., 1995. The promise of public sector-sponsored training programs. Journal of Economic Perspectives 9, 149–168.

Lewis, H.G., 1963. Unionism and Relative Wages. University of Chicago Press, Chicago.

Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Matzkin, R., 1992. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. Econometrica 60 (2), 239–270.

Matzkin, R., 1994. Restrictions of economic theory in nonparametric methods. In: Engle, R., McFodden, D. (Eds.), Handbook of Econometrics, Vol. IV. Elsevier, Amsterdam, pp. 2523–2558.

Moffit, R., 2001. Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice: Comment. Journal of Business and Economic Statistics 19, 20–23.

Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, Vol. IV. Elsevier, Amsterdam.

Neyman, J., 1923. Statistical problems in agricultural experiments. Journal of the Royal Statistical Society (Supplement)2(2), 107–180.

Nowak, L., 1983. A cost effectiveness evaluation of the federal/state vocational rehabilitation program—using a comparison group. The American Economist 27, 23–29.

Poirier, D., 1980. Partial observability in bivariate probit models. Journal of Econometrics 12, 209–217.

Quandt, R., 1972. Methods for estimating switching regressions. Journal of the American Statistical Association 67 (338), 306–310.

Ridder, G., 1986. An event history approach to the evaluation of training, recruitment and employment programmes. Journal of Applied Econometrics 1, 109–126.

Rosenbaum, P., 1995. Observational Studies. Springer, Berlin.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Roy, A., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3, 135–146.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Todd, P., 2001. Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice: Comment. Journal of Business and Economic Statistics 19, 25–27.

Vytlacil, E., 2002. Independence, monotonicity, and latent variable models: an equivalence result. Econometrica 70, 331–341.

Worrall, J., 1988. Benefit and cost models. In: Berkowitz, E. (Ed.), Measuring the Efficiency of Public Programs. Costs and Benefits in Vocational Rehabilitation. Temple University Press, Philadelphia, Pennsylvania, pp. 45–62.