.

Treatment Choice based on semiparametric evaluation methods

Markus Frölich

Discussion paper 2001-16
Department of Economics, University of St. Gallen
Last changes: December 4th, 2001

Abstract:

Most evaluation studies focus on estimating average treatment effects and largely neglect heterogeneous responses in the population. However, the effectiveness of a policy does not only depend on the programmes themselves but also on the selection of participants. Hence for improving overall effectiveness of the policy it would be important to know for each individual which is his optimal programme, not only which programmes are effective on average. In this paper a semiparametric approach to estimating optimal treatment choices is proposed and applied to Swedish rehabilitation programmes. It seems that a much higher re-employment rate could have been achieved through better participant allocation.

Keywords: Optimal treatment choice, statistical decision rules, heterogeneous response, targeting, profiling, policy evaluation, heterogeneous treatment effect, causal effect, matching.
JEL classification: C44, C14, H43, J24, J60

I would like to thank Jeff Smith for numerous comments and careful reading of an earlier draft of this paper. Further, I am grateful to Bo Honoré, Francois Laisney, Michael Lechner, Christof Luchsinger, Ruth Miquel and seminar participants at the Econometric Society World Congress in Seattle, the conference of the Verein für Socialpolitik in Magdeburg and seminars at the Universities of Konstanz and Strasbourg for valuable discussions, comments and suggestions. All errors are mine. This research was supported by the Swiss National Science Foundation, project NSF 4043-058311. Address for correspondence: Markus Frölich, Swiss Institute for International Economics and Applied Economic Research (SIAW), Department of Economics, University of St. Gallen, Dufourstrasse 48, CH-9000 St. Gallen, Switzerland; markus.froelich@unisg.ch, www.markusfroelich.de

1 Introduction

The econometric evaluation of social policies, for instance of active labour market programmes, is becoming more and more common in many countries. Most of these evaluation studies focus on the estimation of *average treatment effects*, i.e. the difference between the expected outcome when participating in a programme A and the expected outcome when participating in a programme B for an individual drawn randomly from the population, see Angrist and Krueger (1999), Heckman, LaLonde, and Smith (1999) and Lechner (2001). ('Non-participation in any programme' is conceptually defined as a separate programme.) Yet, which constructive policy advice can be drawn from these studies for improving the policy under evaluation? How should programmes and participant allocation be re-organized? Or in particular, should programmes with negative treatment effects be completely abolished?

The answer to such questions depends largely on how heterogeneous individuals respond to programme participation. If all individuals have identical treatment effects, non-performing programmes should be eliminated.[1] Yet, if treatment response is heterogeneous, the effectiveness of a policy depends not only on the programmes themselves but also on the allocation of individuals to these programmes. A programme may be beneficial to some individuals but harmful to others, and a negative average treatment effect does not necessarily mean that the programme does not work; it could simply be the case that the wrong participants are assigned (or self-selected) to this programme. Instead of abolishing such programmes it would be more important to improve the allocation of participants to the programmes. With heterogeneous treatment response it is likely that there is no common optimal treatment and that *optimal treatment choice* varies among individuals depending on their characteristics.

Thus studying heterogeneity should be a central issue in policy evaluation. However, as remarked in Manski (2000b, 2001), evaluation studies rarely analyze how treatment response varies among individuals beyond broad subgroup analyses, where average effects are estimated separately for men/women or different age groups. And this although treatment effect heterogeneity seems to be relevant with many policies. For instance, Heckman, Smith, and Clements (1997) tested the constant treatment effect assumption for the JTPA programme (USA) and clearly rejected it. While beneficial to many participants the programme appeared to be harmful to other participants, in the sense that non-participation would have been more advantageous to them. Also Black, Smith, Berger, and Noel (1999) and Manski (2000b) detected effect heterogeneity. See also Heckman and Robb (1985) and Björkland and Moffitt (1987).

Optimal treatment choice has recently been analyzed by Dehejia (1999) and Manski (2000a, b, 2001). (An earlier reference is Wald (1950)). Dehejia (1999) modelled the treatment deci-

[1] At least if market or general equilibrium effects are of limited importance.

sion problem of an individual and, using a Bayesian approach, looked for first-order stochastic dominance between participation and non-participation. Manski (2000a) analyzed the situation where a planner allocates the members of a population to the available programmes. He shows that if the planner aims to maximize mean outcome in the population (utilitarian welfare) the optimal treatment choice is assigning each individual to that programme that promises the largest expected potential outcome *conditional* on the individual's observed characteristics.[2] To analyze feasibility of such a rule when the conditional expected potential outcomes are unknown and must be estimated Manski (2000b) compares two polar statistical treatment rules: the unconditional and the conditional success rule. The unconditional success rule neglects covariate information and assigns all individuals to the (same) treatment with the highest estimated average outcome. The conditional success rule differentiates among individuals according to their covariates and assigns them on basis of estimated expected potential outcomes conditional on their observed covariates. Since in finite samples the conditional outcomes are less precisely estimated than the unconditional mean outcomes, the conditional success rule entails the risk that due to large estimation error individuals might be allocated worse than if they were all assigned to the same programme. However, Manski shows that even at quite small sample sizes the conditional success rule is superior to the unconditional rule if the outcome variable Y is bounded, and suggests using covariate information to its fullest extent.

Hence it is central for the derivation of optimal treatment choices to estimate for each available programme the *conditional expected potential outcomes* given covariates X from a sample of observations on past participants. Two issues need to be tackled for that. First, if the former participants have not been assigned randomly (within an experimental setting) to the programmes, the selection problem must be taken into account, which arises if the (former) par-

[2]This should not be confused with a popular, though fundamentally flawed, alternative approach called *profiling*, which is used in some countries to assign unemployed to active labour market programmes, Australia (OECD 1998), Netherlands (de Koning 1999), or the Worker Profiling and Reemployment Services (DOL 1999, Black, Smith, Berger, and Noel 1999) in the USA, or to assign welfare recipients to welfare-to-work programmes in the USA (Eberts 1998). Profiling is not based on a counterfactual analysis where relative treatment impacts between different programmes (including 'no-participation' as a treatment) are compared. Instead, only the potential outcome in case of 'no-participation' is estimated for each individual and the individuals with the worst estimated outcomes are assigned to the most intensive programmes. E.g. if long-term unemployment is considered as the main risk, the profiling approach would estimate the expected unemployment duration when the individual does not participate in any active labour market programme, but would not take into account the unemployment duration to be expected if the individual would participate in a programme. Thus, profiling rests on the assumption that the individuals with highest long-term unemployment risk are those who gain most from more expensive programmes. However, this argumentation does not necessarily hold, as for instance evidenced in Berger, Black, and Smith (2001), and is particularly fallacious if participation is harmful due to negative treatment effects, as in Bloom et al. (1997), Puhani (1999), Gerfin and Lechner (2000) or Lechner (2000), among others. Furthermore profiling on a single variable is particularly unsuited, if multiple policy goals are pursued, as it does not allow a transparent weighting of these goals.

ticipants in a certain programme are systematically different from the participants in the other programmes (Heckman 1990, Manski 1993). This selection problem can be solved by including in the conditioning vector X all confounding factors, i.e. all variables that affect the selection process as well as the outcome variables, (known as selection on observables or conditional independence assumption, see Heckman and Robb (1985) or Lechner (1999)). Second, with the conditional expected potential outcomes identified from observations on former participants, it remains their estimation. On the one hand, fully nonparametric approaches might be unreliable since the conditional independence assumption usually requires that X contains many variables. On the other hand, parametric regression leads to inconsistent estimates, unless the true conditional expectation function is correctly specified.[3]

This tension between parametric and nonparametric regression motivates the approach developed in this paper: A *semiparametric estimator* for optimal treatment choices, which integrates parametric regression with nonparametric propensity-score-matching estimates of average (unconditional) treatment outcomes. Propensity score matching evades the curse of dimensionality in the estimation of average treatment effects, since regressing on the one-dimensional propensity score is sufficient to remove selection bias, as shown by Rosenbaum and Rubin (1983), and by Imbens (2000) and Lechner (2001) for the evaluation of policies with multiple programmes.[4] The proposed semiparametric estimator rests on the idea that if the parametric specification was correct, parametric and nonparametric estimates should lead to similar average outcomes. On the other hand, large differences indicate a misspecification of the parametric regression plane and the inclusion of the nonparametric average estimates should alleviate large and systematic biases in the estimated conditional expected potential outcomes. This is achieved by incorporating the average differences between the parametric and nonparametric estimates as additional moments into a GMM estimator. This estimator is developed in Section 2 and in Section 3 its finite sample properties are analyzed.

In the second part of the paper, Section 4, treatment effect heterogeneity and optimal treatment choice are analyzed with respect to the re-employment outcome of rehabilitation programmes for long-term sick in Sweden. In previous studies Frölich, Heshmati, and Lechner (2000a, b) found that medical and particularly educational rehabilitation are not successful in fostering re-employment, raising the question whether educational rehabilitation should be abolished completely. It is investigated how individuals should be selected to the programmes and how this optimal allocation deviates from the allocation observed. Further, it is simulated which average re-employment rate could have been achieved if all individuals had been

[3] Fully parametric models are used in the applied statistical treatment assignment systems for active labour market programmes in Canada (Service and Outcome Measurement System, SOMS, (Colpitts 1999)) and the USA (Frontline Decision Support System, FDSS, (Eberts and O'Leary 1999)).

[4] For applications of matching estimators see Angrist (1998), Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Lechner (1999), Gerfin and Lechner (2000) or Jalan and Ravallion (2002).

assigned optimally. It seems that the observed re-employment rate of 46% could have been increased to about 56% by a better participant allocation. Moreover, although educational rehabilitation turns out to be the first-best choice for quite many individuals, abolishing educational rehabilitation and relegating these individuals to their second-best choice decreases the simulated re-employment rate only negligibly. These results are stable for different specifications, and suggest that not only on average educational rehabilitation is harmful to immediate re-employment chances but that furthermore hardly anyone could benefit substantially from educational rehabilitation.[5] Section 5 concludes. Appendix A contains further tables. A supplementary appendix with proofs and additional results is available on www.markusfroelich.de

2 Optimal treatment choice

A useful framework for discussing optimal programme choices and the corresponding decision problem is the potential outcomes framework of Rubin (1974), see also Roy (1951). Suppose the policy under consideration consists of R (exhaustive and mutually exclusive) different programmes, of which each member of an eligible population has to choose exactly one. Usually the policy consists of $R-1$ active treatments, e.g. public employment programmes, on-the-job and classroom training, employment subsidies, and a treatment called 'no-participation' for those who do not participate in any of the active programmes. Let $Y^1, Y^2, .., Y^R$ denote the *potential* outcomes for a certain individual, where Y^1 is the outcome that she would realize if she participated in programme one, Y^2 is the outcome she would realize if she participated in programme two, and so forth. Obviously, only one of these R potential outcomes can be observed ex-post, after participation in the policy.

Most policies are envisaged to pursue multiple, often conflicting goals, measured by different outcome variables and value these goals differently, e.g. re-employment on basis of a permanent contract may be higher valued than re-employment on basis of a fixed-term contract (Brodaty, Crépon, and Fougère 2001). To account for multiple goals the potential outcomes are allowed to be vectors $Y^1, .., Y^R \in \Re^V$ consisting each of V different success measures, e.g. economic, social, health, psychological and cost indicators. Let $u(\cdot) : \Re^V \mapsto \Re$ be a known utility weighting function that weighs these success indicators according to the relative importance attached to the different policy goals. To maximize the overall effectiveness of the policy each individual should be allocated to that programme that yields the largest goals-weighted outcome.[6] Yet, the

[5] Although these results strongly advocate the elimination of educational rehabilitation no hasty policy recommendations should be drawn since the re-employment outcome is measured immediately at the end of the sickness spell and the data contains no follow-up information.

[6] Implementing such an optimal allocation, i.e. enforcing compliance when policy goals and individuals' goals conflict, is another matter and not discussed here.

idiosyncratic potential outcomes can (ex-ante) never be inferred for any particular individual. At most it can be hoped that the distribution of Y^r for this particular individual can be well approximated by the distribution of Y^r, conditional on observed characteristics. Let $x \in \Re^k$ denote a detailed set of individual characteristics. If the policy aims to maximize the expected goals-weighted utility, similar to Manski (2000a, b), the optimal programme r^* for an individual with characteristics x is

$$r^*(x) = \underset{r \in \{1,..,R\}}{\arg\max} E\left[u\left(Y^r\right)|X = x\right].$$

This equals

$$r^*(x) = \underset{r \in \{1,..,R\}}{\arg\max}\ u\left(E[Y^r|X = x]\right), \tag{1}$$

if u is a linear function weighting the different policy goals. Hence, the estimated optimal programme $\hat{r}^*(x)$ for an individual with characteristics x is the programme with the highest estimated goals-weighted conditional potential outcome. Since the joint distribution of the potential outcome vectors $Y^1, .., Y^R$ is not identified the estimation of the potential outcomes $E[Y^r|X = x]$ can proceed separately for each $r \in \{1, .., R\}$.

However, this procedure does not take into account the estimation variability of the estimates $\hat{Y}^r(x) = \hat{E}[Y^r|X = x]$. In practice it would be relevant to know how strong the evidence is that a certain programme is the optimal one, since very noisy estimates of the conditional outcomes might pretend treatment choice heterogeneity, even if there is a common optimal treatment. To assess such a situation multiple-comparisons-with-the-best (MCB) techniques have been developed (Hsu 1996, Horrace and Schmidt 2000), which provide a ordering of significantly different programmes and test the hypothesis that programme r^* is the best treatment at a confidence level α, i.e. that

$$P\left(\underset{r \in \{1,..,R\}}{\arg\max}\ u\left(\hat{Y}^r(x)\right) = r^*\right) \geq 1 - \alpha. \tag{2}$$

However MCB methods are usually designed for consistent parametric estimators and thus incorporate only variance but not bias. Further, with more complex estimators the variance component is usually estimated on basis of asymptotic approximations which may be less precise in finite samples. Instead, I propose simulating the joint distribution of $\hat{Y}^1(x), \hat{Y}^2(x), .., \hat{Y}^R(x)$ by bootstrap simulation and approximate such for a particular individual the probability that a certain programme is preferable to all others. For larger numbers of available programmes R it will often be the case that no programme dominates all other with high probability, but that a semi-ordering into best, intermediate, and worst programmes is possible. The subset of best programmes jointly dominates all other programmes with high probability, but among these best programmes no statistically significant ordering is possible. If evidence is weak, the individual might be allocated to a programme on basis of other considerations (e.g. waiting time for treatment or as a fill-in person).

If the utility function is unknown, for instance if individuals self-select to the programmes on basis of their own decision-making process, it may still be possible in some cases to derive dominance relationships between programmes e.g. for a subset of the outcome measures. Additionally, all estimated potential outcomes should be provided to the decision makers as additional information to improve their decision making (Mohr 1999).

2.1 Conditional expected potential outcomes

As discussed above, central for estimating optimal treatment choices are the conditional expected potential outcomes $E[Y^r|X]$, which can only be estimated from data on former participants. Let $\{(X_i, D_i, Y_i)\}_{i=1}^n$ be a sample from a previously treated population, where $D_i \in \{1, .., R\}$ indicates in which programme individual i had participated and $Y_i = Y_i^{D_i}$ denotes the outcome observed (after participation). However, since Y^r is observed only for the participants in r, only $E[Y^r|X, D = r]$ is identified from this sample, but not $E[Y^r|X]$. If the former participants have been selected to the programmes in a non-random way, the participants in a particular programme might be systematically different from the participants in other programmes. Then it is likely that $E[Y^r|X, D = r]$ differs from the *counterfactual* outcome $E[Y^r|X, D \neq r]$, i.e. the outcome that individuals who participated in other programmes would have realized if they had participated in programme r instead.

One approach to solve this fundamental problem of evaluation (see e.g. Manski 1993) is to include in X all variables that affected the potential outcomes as well as the programme assignment under the former selection process. (This, of course, requires that all these variables are observed in the data set.)[7] Then the potential outcome Y^r is *conditionally independent* of whether an individual selected into programme r or not:

$$Y^r \perp\!\!\!\perp \mathbf{1}(D = r) \,|X \qquad\qquad \forall r \in \{1, .., R\}, \tag{3}$$

where $\perp\!\!\!\perp$ denotes statistical independence and $1(\cdot)$ is the indicator function.[8] This assumption implies $E[Y^r|X = x] = E[Y^r|X = x, D = r]$ and identifies the conditional potential outcome, provided that there is a positive probability that an individual with characteristics x is observed participating in programme r. Let $p^r(x) \equiv P(D = r|X = x)$ denote the probability for an individual with characteristics $X = x$ to have been selected into programme r, and let $S^r =$

[7] If the available data are not sufficiently informative, but contain a local instrumental variable IV regression methods might be used (Imbens and Angrist 1994, Angrist, Imbens, and Rubin 1996, Heckman and Vytlacil 1999). Alternatively, no-assumption bounds as in Manski (1997, 2000b) could be derived. However, it remains open the question how to proceed if these bounds are non-informative, since a treatment choice must be made in any case. Obviously, if experimental data were available, as in Manski (2000b), the analysis simplifies greatly.

[8] With other words, treatment selection depended on the potential outcomes only to the extent to which they could be anticipated on basis of the exogenous characteristics X, but not on an anticipation based on unobserved characteristics.

$\{x : p^r(x) > 0\}$ denote the support of X among the participants in programme r.[9] Then $E[Y^r|X = x]$ is identified by the sample of former participants for all $x \in S^r$.

With $E[Y^r|X]$ nonparametrically identified estimation becomes the crucial issue. Conditional expectation functions are often estimated by parametric regression. Assuming that the conditional expectation function is known to belong to a class of functions $\varphi^r(x;\theta^r)$, e.g. linear, with unknown but finite coefficient vector θ^r, the coefficients θ^r can be estimated (by Maximum Likelihood or GMM) and the conditional expected outcomes can immediately be predicted at any x. However, these estimates are usually inconsistent, except if the true conditional expectation function is really contained in the specified class $\varphi^r(\cdot;\cdot)$. Thus, nonparametric regression would be more appealing. But, for the conditional independence assumption (3) to be valid often many variables have to be included in X, rendering nonparametric regression difficult, see for instance Härdle (1991, Ch.10).

Because of this dimensionality problem most evaluation studies have refrained from estimating conditional potential outcomes $E[Y^r|X]$ and concentrated on nonparametric estimation of average (unconditional) potential outcomes $E[Y^r]$ or average counterfactual outcomes $E[Y^r|D \neq r]$, for calculating average treatment effects. Their estimation is eased by the balancing property of the propensity score, which states that independence conditional on X (3) also implies independence conditional on the one-dimensional participation probability $p^r(X)$:

$$E[Y^r|p^r(X) = p] = E[Y^r|p^r(X) = p, D = r], \tag{4}$$

as shown by Rosenbaum and Rubin (1983), and by Imbens (2000) and Lechner (2001) for the evaluation of multiple programmes ($R > 2$). By the law of iterated expectations the average counterfactual outcome is identified for the population contained in the support S^r by[10]

$$E_{S_r}[Y^r|D \neq r] = E_{S_r}\left[E[Y^r|p^r(X), D \neq r] \;|D \neq r\right] = E_{S_r}\left[E[Y^r|p^r(X), D = r] \;|D \neq r\right] \tag{5}$$

where $E[Y^r|p^r, D = r]$ can be estimated by *one-dimensional* nonparametric regression, e.g. nearest neighbour matching or local polynomial regression (Fan 1992, Hastie and Loader 1992). Thus, conditioning on the one-dimensional participation probability p^r suffices to eliminate selection bias. Let $\hat{p}^r$ and $\hat{m}^r$ be preliminary estimates of the participation probability $p^r(x)$ and the regression curve $m^r(p) = E[Y^r|p^r(X) = p]$ for $p > 0$, respectively, and $\hat{S}^r = \{x : \hat{p}^r(x) > 0\}$ an estimate of the support region, then the average counterfactual outcome $E_{S_r}[Y^r|D \neq r]$ can

[9] This is equivalent to $S^r = \{x : f_{X|D=r}(x) > 0\}$ where $f_{X|D=r}$ is the density of X in the subpopulation of participants in treatment r, because $p^r(x) = f_{X|D=r}(x)P(D = r)/f_X(x)$ and $p^r(x)$ is undefined where f_X, the density of X in the population, is zero.

[10] Since the conditional outcome $E[Y^r|X]$ is not identified outside the support S^r, the population expectation $E[Y^r]$ is not a useful concept.

be estimated by a generalized matching estimator as

$$E_{S_r}[\widehat{Y^r|D \neq r}] = \frac{\sum \hat{m}^r\left(\hat{p}_i^r\right) \cdot 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)}{\sum 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)}, \tag{6}$$

where $\hat{p}_i^r = \hat{p}^r(X_i)$. (If Y^r is a vector then all operations with respect to Y^r are defined as *element-wise* in Y^r.) This estimator essentially re-weights the expected outcome conditional on p^r by the different distributions of p^r among the participants and the non-participants, i.e. it centers the expectation of Y^r at the distribution of the non-participants. Heckman, Ichimura, and Todd (1998) analyzed the generalized matching estimator with m^r estimated by local polynomial regression and established $\sqrt{n}$-consistency and asymptotic normality. Frölich (2000) analyzed the small-sample properties of various generalized matching estimators and found that *SG matching*, based on a modified local linear version of Seifert and Gasser (1996, 2000), performed best and significantly better than pair-matching and local linear regression.

This tension between parametric estimation of conditional potential outcomes $E[Y^r|X]$, which is inconsistent if misspecified, and nonparametric estimation of average potential outcomes $E[Y^r]$, that are largely uninformative about heterogeneous treatment response,[11] motivates one of the main proposals of this paper: to combine a parametric specification for the conditional potential outcomes with nonparametrically estimated average potential outcomes. The nonparametric estimates would contribute to avoiding large and systematic biases among the non-participants, which are neglected by the parametric estimator.

Suppose that the conditional expectation functions are parametrically specified by (vector-valued) functions $\varphi^r(x, \theta^r)$ with coefficient vectors θ^r of dimension k as

$$E[Y^r|X = x] \doteq \varphi^r(x, \theta^r) \qquad \forall r \in \{1, .., R\}. \tag{7}$$

If the expectation functions were correctly specified, there would exist true coefficient vectors θ_0^r such that the equations (7) would hold with equality for all x. With consistent estimates $\hat{\theta}^r$ the average counterfactual outcome would consistently be estimated by the parametric model as

$$E_{S_r}[\widehat{Y^r|D \neq r}] = \frac{\sum \varphi^r(X_i, \hat{\theta}^r) \cdot 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)}{\sum 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)} \tag{8}$$

and should thus be similar to the nonparametric estimate (6).

On the other hand, if the functions φ^r are misspecified, the predictions of $E[Y^r|X = x]$ might be severely biased. This might occur particularly in regions with few participants in programme r but many non-participants, since the parametric estimator of θ^r takes only account of the observations for whom Y^r is observed. I.e. the parametric estimator is centered at the distribution of the participants and neglects the characteristics of the non-participants. But as

[11] Also the conditional expectation function $E[Y^r|p^r(X)]$ is not very informative for choosing a treatment for a particular individual.

the participants in a certain programme r are often rather different from the non-participants, predictions outside the subpopulation of participants can be poor. To avoid large and systematic differences between the predicted potential outcomes for the non-participants according to the parametric model and the nonparametric estimates, the *difference* between (8) and (6)

$$\frac{\sum \left(\varphi^r(X_i,\theta^r) - \hat{m}^r\left(\hat{p}_i^r\right)\right) \cdot 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)}{\sum 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)} \tag{9}$$

should be small, i.e. the 'out-of-sample' predictions to the non-participants should lead to similar results. Otherwise, a large deviation indicates a serious misspecification of the parametric model.

This should also hold analogously for any subpopulation defined on the X characteristics. Let $\Lambda(x)$ be a $L \times 1$ vector-valued indicator function defining L different subpopulations. For instance the 3 populations: *all*, *men*, *age 40 to 50 years* would be defined by

$$\Lambda(x) = \begin{pmatrix} 1 \\ 1(x_{\text{gender}} = \text{male}) \\ 1(x_{\text{age}} \in [40,50]) \end{pmatrix}, \tag{10}$$

where x_{gender} and x_{age} refer to the respective elements of the x characteristics vector. In analogy to (9) θ^r should be chosen in a way to keep the 'out-of-sample' prediction differences between parametric model and nonparametric estimates

$$\frac{\sum \left(\Lambda(X_i) \otimes \varphi^r(X_i,\theta^r) - \hat{\mathbf{m}}_{VL}^r(\hat{p}_i^r)\right) \cdot 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)}{\sum 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right)} \tag{11}$$

small in all subpopulations, where $\otimes$ is the Kronecker product operator. $\hat{\mathbf{m}}_{VL}^r(\cdot)$ is the $VL \times 1$ column vector of all stacked nonparametric estimators of $E[Y^r|p^r]$ for all populations, *multiplied* with the population indicator function. I.e. the first V elements of $\hat{\mathbf{m}}_{VL}^r$ represent the estimators for all V outcomes measures for subpopulation 1, the following V elements correspond to subpopulation 2 and so forth. More precisely, let $\hat{m}_{vl}^r(p)$ for $p > 0$ be an estimator of the expectation $E[Y_v^r|p^r(X) = p, \Lambda_l(X) = 1]$, i.e. the expectation of the v-th variable of the potential outcome vector Y^r conditional on the participation probability p^r in the l-th subpopulation defined by $\Lambda_l(X)$. Let $\hat{m}_l^r(\cdot) = (\hat{m}_{1l}^r(\cdot), .., \hat{m}_{vl}^r(\cdot), .., \hat{m}_{Vl}^r(\cdot))'$ be the element-wise-defined estimator of the potential outcome vector Y^r in the population l, i.e. of $E[Y^r|p^r(X) = p, \Lambda_l(X) = 1]$. Stacking these estimators for the L subpopulations and multiplying element-wise with the population indicator function gives

$$\hat{\mathbf{m}}_{VL}^r(\hat{p}^r(X_i)) = \left(\hat{m}_1^{r\prime}\left(\hat{p}^r\left(X_i\right)\right) \cdot \Lambda_1\left(X_i\right), .., \hat{m}_l^{r\prime}\left(\hat{p}^r\left(X_i\right)\right) \cdot \Lambda_l\left(X_i\right), .., \hat{m}_L^{r\prime}\left(\hat{p}^r\left(X_i\right)\right) \cdot \Lambda_L\left(X_i\right)\right)'.$$

Including additional subpopulations in (11) anchors the parametric model more closely at the nonparametric estimates and thus controls deviations more finely. On the other hand, if smaller subpopulations are included their nonparametric estimates will be less precise and their additional value as anchor for the parametric model will be limited. Determining the optimal number and size of populations included depends itself on the degree of heterogenous response and is left for future research.

2.2 Semiparametric GMM estimator

For coalescing the parametric and nonparametric elements the GMM framework seems ideally suited. The parametric specification (7) implies $E[A^r(X) \cdot (Y^r - \varphi^r(X, \theta_0^r))] = 0$, with $A^r(X)$ an instrument matrix, if the parametric specification φ^r is correct, i.e. a true coefficient vector θ_0^r exists. Choosing A^r to be a $k \times V$ matrix would lead to a just identified parametric GMM estimator, as θ_0^r contains k coefficients. To avoid estimates of θ^r that lead to large biases among the non-participants in case of misspecification, these k moments emanating from the parametric model can be augmented by a second set of VL moments according to the distance vector (11). If the parametric specification is correct both the parametric and the nonparametric estimator of the average counterfactual outcome $E_{S_r}[Y^r|D \neq r]$ are consistent and the difference (11) converges to zero. This leads to the moment vector

$$g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{p}^r) = \frac{1}{n}\sum_i g_i^r = \frac{1}{n}\sum_i \begin{pmatrix} A^r(X_i) \cdot (Y_i - \varphi^r(X_i, \theta^r)) \cdot 1(D_i = r) \\ \left(\Lambda(X_i) \otimes \varphi^r(X_i, \theta^r) - \hat{\mathbf{m}}_{VL}^r(\hat{p}_i^r)\right) \cdot 1(D_i \neq r) \cdot 1\left(\hat{p}_i^r > 0\right) \end{pmatrix} \tag{12}$$

of length $k + VL$. The first k moments are evaluated for the observations with $D_i = r$, since only for the participants Y^r can be observed. The lower part of the moment vector compares the 'out-of-sample' prediction for the non-participants with the nonparametric estimate.[12] The coefficients θ^r are estimated by GMM as the solution to

$$\hat{\theta}_n^r = \arg\min_{\theta^r} g_n^{r\prime} W^r g_n^r, \tag{13}$$

where W^r is a positive semidefinite matrix that attaches different weights to the moments. Attaching zero weight to the VL nonparametric moments would lead to the fully parametric estimator discussed above.

If the parametric specification is correct and the nonparametric estimator $\hat{m}^r(\hat{p}^r)$ of $E[Y^r|p^r]$ is asymptotically linear with trimming, as for instance local polynomial regression (Heckman, Ichimura, and Todd 1998), then the coefficient estimates $\hat{\theta}^r$ are consistent and $\sqrt{n}$-asymptotically normal with approximate variance

$$\frac{1}{n}(G^{r\prime}W^rG^r)^{-1}G^{r\prime}W^rE[J^rJ^{r\prime}]W^rG^r(G^{r\prime}W^rG^r)^{-1}, \tag{14}$$

[12] Obviously, the second set of moments could be further augmented by including the average out-of-sample prediction differences separately for the participants in programme 1, programme 2 and so forth, instead of aggregating all non-participants through $1(D_i \neq r)$. But, some of these moments might then rest on only a small number of effective observations if the number of participants is unevenly distributed among the available programmes. For instance, in the application in Section 4 the largest participant group consists of about 3500 observations whereas the smallest contains only 360 observations. Instead, it seems more flexible to generate additional moments by defining further subpopulations $\Lambda(X_i)$ on basis of the X characteristics, to ensure that the subpopulations are not too small.

where $G^r = E\left[\frac{\partial g_n^r(\theta_0^r, \hat{\mathbf{m}}_{VL}^r, \hat{p}^r)}{\partial \theta'}\right]$ is the expected gradient. Further

$$J^r = g^r(Y, D, X, \theta_0^r, \mathbf{m}_{VL}^r)$$
$$- \begin{pmatrix} \mathbf{0}_k \\ \lambda_{1,r}^{-1} \cdot E[\Psi_{11,m}^r(Y, D, X; X_2) 1(D_2 \neq r) | Y, D, X] + E[\Psi_{11,p}^r(Y, D, X; X_2) 1(D_2 \neq r) | Y, D, X] \\ \vdots \\ \lambda_{L,r}^{-1} \cdot E[\Psi_{VL,m}^r(Y, D, X; X_2) 1(D_2 \neq r) | Y, D, X] + E[\Psi_{VL,p}^r(Y, D, X; X_2) 1(D_2 \neq r) | Y, D, X] \end{pmatrix},$$

where the expectation operator is with respect to X_2 and D_2, and $\lambda_{l,r} = \lim_{n \to \infty} \frac{n_{l,r}}{n}$, with $n_{l,r}$ the number of participants in treatment r belonging to subpopulation l. The influence functions $\Psi_{vl,p}^r$ and $\Psi_{vl,m}^r$ take account of the variance due to the preliminary estimation of the participation probabilities $p^r(\cdot)$ and the regression curve $m^r(\cdot)$, respectively.[13]

Furthermore the statistic

$$n \cdot g_n^{r\prime} \hat{\Omega}^r g_n^r \xrightarrow{d} \chi^2_{(VL)}, \tag{15}$$

with $\hat{\Omega}^r$ a consistent estimate of $[EJ^r J^{r\prime}]^{-1}$, is asymptotically χ^2 distributed with number of freedoms equal to the number of overidentifying restrictions VL. Hence the J-test of overidentifying restrictions (Hansen 1982) can be used to test whether the parametric specification is correct.

If, on the other hand, the parametric specification (7) is incorrect the GMM estimator tries to fit the parametric plane among the participants and to minimize the discrepancy between the parametric out-of-sample prediction and the nonparametric estimates. In this situation the nonparametric estimates should help to avoid systematically large biases for the non-participants. A general proof of this property, however, seems difficult. Although it is obvious that attaching positive weight to the semiparametric moments in (13) leads asymptotically to a smaller average bias among the non-participants, this relationship is unclear in finite samples and also with respect to average squared bias or average squared error. Therefore the finite sample properties under correct and incorrect specification are examined below.

3 Monte Carlo simulation

In a small Monte Carlo experiment the precision in estimating conditional expected potential outcomes $E[Y^1|X = x]$ is compared for three estimators: The fully parametric estimator, the GMM estimator with identity weighting matrix and the GMM estimator with $\hat{\Omega}^r$ weighting matrix. The parametric estimator corresponds to the solution of (13) with a weighting matrix W^r that assigns zero weights to the VL nonparametric moments. The first GMM estimator

[13] Proofs and expressions for $\Psi_{vl,p}^r$,$\Psi_{vl,m}^r$ are given in the supplementary appendix on www.markusfroelich.de

employs an identity matrix for W^r, and the second GMM estimator uses the first GMM estimates to estimate the inverse of the covariance matrix of the moment vector by $\hat{\Omega}^r = [\hat{E}J^r J^{r\prime}]^{-1}$, which is then employed as the weighting matrix in (13). In a (correctly specified) parametric setup the second GMM estimator would represent the efficient GMM estimator (Hansen 1982). However, in the nonparametric approach considered in this paper the second GMM estimator is not necessarily superior to the first GMM estimator, since the 'efficient' weighting by $[EJ^r J^{r\prime}]^{-1}$ takes only notice of variance but not of the bias of the parametric specification. This leads to a weighting matrix $\hat{\Omega}^r$ which assigns most of the weight to the k parametric moments and little to the nonparametric moments, since the variance of the nonparametric estimates is much higher compared to the parametric moments. Accordingly the GMM estimator with the weighting matrix $\hat{\Omega}^r$ is governed by the parametric moments and its coefficient estimates are usually more similar to those of the fully parametric estimator. However, the uncertainty which stems from not knowing the true form of the conditional expectation function is not incorporated in these weights such that such robustness-to-misspecification considerations are neglected in the weighting matrix $\hat{\Omega}^r$. Thus, in case of serious misspecification the second GMM estimator might pay too little attention to the nonparametric estimates.

The Monte Carlo simulations proceed by repeatedly drawing samples $\{(X_i, D_i, Y_i^1 D_i)\}_{i=1}^n$, estimating the coefficients θ^1 and computing (out-of-sample) mean squared prediction error (MSE) by comparing the estimated potential outcome $\hat{E}[Y^1|X = x]$ with the true conditional expected outcome for 10^5 different values of x (drawn from the same population as the sample). $D_i \in \{0, 1\}$ is binary, and the potential outcome Y_i^1 is one-dimensional (V=1) and is observed only if $D_i = 1$. Restricting D_i to be binary leads to no loss of generality, since the GMM estimator (12) distinguishes only between observations with $D_i = r$ and $D_i \neq r$. In this section the superscript r is henceforth suppressed.

The parametric estimator computes θ by least squares. The first GMM estimator uses the least squares estimates as starting values and retains these if optimization of the GMM function (13) fails. With these GMM estimates the asymptotic covariance matrix $E[JJ']$ of the moment vector g_i is calculated and the second GMM estimator is computed with the inverse of this covariance matrix as weighting matrix. As preliminary estimates the participation probabilities $\hat{p}_i$ are estimated by Probit and the regression curves $m(p) = E[Y|p(X) = p]$ are estimated for the various subpopulations by the local linear regression variant of Seifert and Gasser (1996, 2000) with Epanechnikov kernel and bandwidth chosen by cross-validation.

The X characteristics consist of 3 explanatory variables (X_{i1}, X_{i2}, X_{i3}) drawn from the (non-symmetric) $\chi^2_{(2)}, \chi^2_{(3)}, \chi^2_{(4)}$ distribution and divided by 2,3,4, respectively, to standardize their mean. A value of 0.5 is added such that their minimum value is 0.5 (and mean 1.5). The observations X_i are assigned to the treatment and control group according to the selection rule $D_i = 1(X_{i1} + X_{i2} + X_{i3} + \varepsilon_i > 4.5)$ with ε standard normal distributed. About 46% of the population are assigned to the treatment group (D=1).

The Y_i data are generated according to one of three different Y-models with ξ a standard normal error term:

$$\begin{aligned}
&\text{Ymodel 1} \quad Y_i = X_{i1}^2 + X_{i2}^2 + X_{i3}^2 + \xi_i \\
&\text{Ymodel 2} \quad Y_i = \sqrt{X_{i1} - 0.5} + 2\sqrt{X_{i2} - 0.5} - \sqrt{X_{i3} - 0.5} + \xi_i \\
&\text{Ymodel 3} \quad Y_i = X_{i1}X_{i2} + X_{i1}X_{i3} + X_{i2}X_{i3} + \xi_i.
\end{aligned}$$

Four different specifications for the parametric component $\varphi(x, \theta)$ are examined, which are all linear models and vary in their regressor-set included:

Specification	Number of regressors	Regressors
φ_0	$k = 4$	*const*, X_{i1}, X_{i2}, X_{i3}
φ_1	$k = 4$	*const*, X_{i1}^2, X_{i2}^2, X_{i3}^2
φ_2	$k = 4$	*const*, $\sqrt{X_{i1} - 0.5}$, $\sqrt{X_{i2} - 0.5}$, $\sqrt{X_{i3} - 0.5}$
φ_3	$k = 7$	*const*, X_{i1}, X_{i2}, X_{i3}, $X_{i1}X_{i2}$, $X_{i1}X_{i3}$, $X_{i2}X_{i3}$.

Hence the specification φ_0 is incorrect for all Ymodels, φ_1 is correct only for Ymodel 1, φ_2 is correct only for Ymodel 2, and φ_3 is correct only for Ymodel 3.

The GMM estimators are examined for various numbers of subpopulations L included. The GMM estimator with one nonparametric moment (L=1) includes the difference between the parametric and the nonparametric average counterfactual outcome in the entire population. The GMM estimator with four nonparametric moments (L=4) includes additionally the moments for the three subpopulations defined by $X_1 < 1.5$, $X_2 < 1.5$, and $X_3 < 1.5$, respectively. The estimator with L=7 nonparametric moments adds the three subpopulations: $\{X_1 < 1.5 \wedge X_2 < 1.5\}$, $\{X_1 < 1.5 \wedge X_3 < 1.5\}$ and $\{X_2 < 1.5 \wedge X_3 < 1.5\}$. The estimator with L=10 nonparametric moments includes further the subpopulations defined by $X_1 < 1$, $X_2 < 1$, and $X_3 < 1$, respectively, and finally the estimator with L=14 moments appends the subpopulations $X_1 > 2$, $X_2 > 2$, $X_3 > 2$, and $\{X_1 < 1.5 \wedge X_2 < 1.5 \wedge X_3 < 1.5\}$. Whereas the first four populations are large and cover each at least 60% of the population, the subsequent subpopulations become smaller (populations five to seven cover each about 37% and subpopulations eight to ten each about 30% of the population) and the last four subpopulations cover each only about 20% of the population. Thus, not only does the number of overidentifying moment restrictions grow with increasing L but also the precision of the estimated nonparametric averages decreases since the additional populations are smaller and render nonparametric estimation more difficult. Hence, it is to expect in this setup that the GMM estimators with a large number of moments should be becoming imprecise.

The expected outcomes vary considerably among these subpopulations. Whereas in Ymodel 1 the expected potential outcome EY^1 in the population is 13.1 for the participants and 5.3 for the non-participants, the outcome difference between participants and non-participants can be as large as 8.2 (for the populations ten and eleven) and as small as 0.8 (for population fourteen).

Similar heterogeneity occurs for Ymodel 2 and 3. For instance, in Ymodel 2 the expected outcome for the participants is usually larger than for the non-participants, but this relationship is reversed in subpopulation five. In Ymodel 2 the expected outcomes for participants and non-participants are 2.2 and 1.5, respectively, and in Ymodel 3 these figures are 9.6 and 4.3.

In Table 3.1 the mean squared error of the fully parametric and both GMM estimators is given for the (incorrect) specifications (φ_0,φ_2,φ_3) when the Y_i observations are generated by Ymodel 1. The left part of the table provides the results for sample size 500, the right part for sample size 2000. The row L=0 contains the MSE of the parametric estimator (with no nonparametric moment conditions), and the results for the GMM estimators are given for various number of included moments (L=1,4,7,10,14), with the MSE of the first GMM estimator in the left column and the MSE of the second GMM estimator in the right column. It is seen that the GMM estimators have always lower MSE than the fully parametric estimator, and that the first GMM estimator generally performs somewhat better than the second. The MSE of the second GMM estimator usually increases with the number of moments included, whereas the first GMM estimator is less susceptible to the number of moments. Regarding sample size, the MSE of the parametric estimator decreases only little when the sample size is quadrupled, while the GMM estimators become relatively more precise. Generally, the relative reductions in MSE vis-a-vis the parametric estimator are between 20-36% for the first GMM estimator and 16-34% for the second GMM estimator at sample size 500 and 20-44% (16-39%) at sample size 2000.

Table 3.1: Mean squared error with incorrect parametric specification for Ymodel 1

Ymodel 1	n=500						n=2000					
	φ_0		φ_2		φ_3		φ_0		φ_2		φ_3	
L=0	9.75		22.14		13.23		9.53		21.21		13.12	
L=1	7.08	7.52	17.63	17.21	8.62	8.70	6.58	7.24	17.05	16.76	7.55	8.03
L=4	7.27	7.74	17.27	17.35	8.44	9.23	6.69	7.26	16.75	17.22	7.41	8.26
L=7	7.38	7.91	17.24	17.45	8.46	9.35	6.77	7.34	16.73	17.27	7.45	8.37
L=10	7.48	8.14	17.21	18.15	8.58	9.90	6.82	7.39	16.71	17.88	7.51	8.75
L=14	7.49	8.22	17.20	18.08	8.50	9.52	6.84	7.63	16.72	17.81	7.52	8.42

Note: MSE for various parametric specifications (φ_0,φ_2,φ_3) for the fully parametric estimator (L=0), and for the first GMM estimator (left column) and the second GMM estimator (right column) for different numbers of nonparametric moments (L=1,4,7,10,14). The total number of moments is $L+k$ where k=4 in specifications $\varphi_0, \varphi_1, \varphi_2$ and k=7 in φ_3. The true data generating process corresponds to Y-model 1. 1000 replications.

The results according to Ymodel 2 are given in Table 3.2. For sample size 2000 the results are similar to those of Table 3.1. The GMM estimators are generally more precise than the parametric estimator, with the first GMM estimator usually being preferable to the second. However, in specification φ_3 the first GMM estimator becomes slightly inferior to the parametric estimator when the number of nonparametric population moments L exceeds the number of

X regressors, which is $k = 7$ for specification φ_3. For sample size 500 the results are mixed. Whereas the first GMM estimator is clearly the best with specification φ_1, it is usually somewhat worse with specification φ_0. In specification φ_3 the first GMM estimator worsens considerably with growing sample size and the dominant estimator is the second GMM estimator. However, the first GMM estimator's large MSE when the number of moments is large reduces by half when the sample size is increased to 2000, whereas it decreases only slightly for the other two estimators. At sample size 2000 the precision gains of the GMM estimators relative to the parametric estimator are between -3 to 19% for the first and 3 to 17% for the second GMM estimator.

Table 3.2: Mean squared error with incorrect parametric specification for Ymodel 2

Ymodel 2	n=500						n=2000					
	φ_0		φ_1		φ_3		φ_0		φ_1		φ_3	
L=0	9.72		36.90		12.30		8.27		35.50		9.92	
L=1	10.33	9.40	32.68	33.92	11.56	11.14	7.77	8.05	30.05	31.92	8.38	9.15
L=4	10.24	10.09	33.15	34.58	13.28	11.44	7.53	7.81	30.46	31.89	8.09	8.26
L=7	10.74	10.23	33.49	34.88	17.42	11.63	7.55	7.90	30.62	31.90	8.97	8.33
L=10	10.97	10.38	33.60	35.09	20.89	11.82	7.47	7.94	30.79	32.09	10.35	8.24
L=14	11.09	10.30	33.73	35.19	20.51	11.93	7.50	7.90	30.84	32.00	10.26	8.23

Note: See note below Table 3.1. All figures multiplied by 100. True data generating process is Y-model 2.

Table 3.3: Mean squared error with incorrect parametric specification for Ymodel 3

Ymodel 3	n=500						n=2000					
	φ_0		φ_1		φ_2		φ_0		φ_1		φ_2	
L=0	2.45		4.49		6.39		2.39		4.32		6.25	
L=1	1.65	1.74	4.20	4.26	3.05	3.20	1.54	1.64	3.96	4.19	2.93	3.07
L=4	1.75	1.84	4.20	4.27	3.14	3.26	1.59	1.63	3.99	4.27	2.96	3.05
L=7	1.82	2.01	4.20	4.33	3.24	3.39	1.64	1.76	3.98	4.26	3.02	3.13
L=10	1.84	2.06	4.20	4.40	3.28	3.52	1.65	1.79	4.00	4.32	3.04	3.23
L=14	1.86	2.08	4.20	4.42	3.30	3.52	1.67	1.79	4.00	4.39	3.06	3.23

Note: See note below Table 3.1. True data generating process is Y-model 3.

The results for Ymodel 3 in Table 3.3 indicate a clear superiority of the GMM estimators for all specifications φ_0, φ_1 and φ_2, with the first GMM estimator dominating in all cases. The mean squared error of the GMM estimators is usually much lower than the MSE of the parametric estimator and the efficiency gains are in some cases larger than 50%, particularly when only few nonparametric population moments are included. The best results are usually obtained with $L \leq k$, i.e. less (or equal) overidentifying moments than the number of regressors.

Taken together, these results suggest that the inclusion of nonparametric population mo-

ments can increase the precision in estimating conditional expected potential outcomes substantially (even reduce MSE up to half) when the true form of the expected potential outcomes is unknown.

In Table 3.4 the properties of the GMM estimator in case of correct specification of the parametric model are examined, i.e. Ymodel 1 with specification φ_1, Ymodel 2 with φ_2 and Ymodel 3 with φ_3. In the upper part the MSE for the parametric and the GMM estimators is given. It is seen, that although the mean squared error of the GMM estimators is relatively much higher than that of the parametric estimator, the absolute magnitude in precision loss is negligible when compared to the absolute efficiency gains seen in Tables 3.1 to 3.3, particularly for Ymodels 1 and 3. If one knew the correct specification with (almost) certainty this would be of high concern. Otherwise these efficiency losses of the GMM estimators are outweighed by the improved robustness in case of misspecification.

Further, the size of the test for overidentifying restrictions is analyzed, which tests whether the parametric model is correctly specified. Since the J-test (15) is known to tend to overreject in many situations (Altonji and Segal 1996, Burnside and Eichenbaum 1996, Hall and Horowitz 1996, Imbens, Spady, and Johnson 1998), alternatively also a Lagrange Multiplier (LM) test proposed in Imbens, Spady, and Johnson (1998) is examined, which is also asymptotically $\chi^2_{(VL)}$ distributed. Imbens, Spady, and Johnson (1998) analyzed various alternative test statistics of which the LM test

$$LM = \hat{\lambda}' \left(\sum \hat{\pi}_i g_i g_i' \right) \left[\sum \hat{\pi}_i^2 g_i g_i' \right]^{-1} \left(\sum \hat{\pi}_i g_i g_i' \right) \hat{\lambda} \xrightarrow{d} \chi^2_{(VL)} \tag{16}$$

performed best in their Monte Carlo simulations, where $\hat{\pi}_i$ are estimated empirical likelihood (or exponential tilting) probabilities and $\hat{\lambda}$ estimated Lagrange multipliers. A convenient alternative to compute the empirical probability weights $\hat{\pi}_i$ provides the estimator of Back and Brown (1993)

$$\hat{\pi}_i = \frac{1}{n} \frac{1 - g_n \hat{\Omega} g_i}{1 - g_n \hat{\Omega} g_n},$$

which comes in a closed form solution and is semiparametrically efficient in estimating the empirical distribution function, see Brown and Newey (2001). With this estimator of the empirical probabilities the lagrange multiplier is

$$\hat{\lambda} = -\hat{\Omega} g_n,$$

where $\hat{\Omega} = [\hat{E} J J']^{-1}$ is the inverse of the estimated covariance matrix of the moment vector, see also Brown, Newey, and May (2001) or Inkmann (2001, p.98 ff.). The J-test (15) and the LM-test (16) are computed for the first and for the second GMM estimator, with g_i and $\hat{\Omega}$ estimated at their respective estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, and are compared to their nominal size of 5% and 10%, respectively.

Table 3.4: Mean squared error and J-tests when specification φ is correct

		n=500						n=2000					
		Ymodel 1		Ymodel 2		Ymodel 3		Ymodel 1		Ymodel 2		Ymodel 3	
L=0		0.01		2.01*		0.03		0.00		0.52*		0.01	
L=1		0.09	0.02	3.48	2.28	0.09	0.03	0.01	0.00	1.14	0.55	0.02	0.01
L=4		0.12	0.02	6.24	2.99	0.14	0.05	0.02	0.00	1.80	0.79	0.02	0.01
L=7		0.14	0.02	9.32	3.17	0.17	0.07	0.02	0.00	2.60	0.90	0.04	0.01
L=10		0.16	0.03	10.87	3.50	0.20	0.09	0.03	0.00	3.26	1.07	0.04	0.01
L=14		0.16	0.04	10.44	3.47	0.19	0.09	0.03	0.00	3.11	1.03	0.04	0.02
Tests		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
L=1:	J_1	71.1	65.6	26.8	20.3	61.4	54.2	57.1	49.8	38.9	32.1	51.3	43.0
	LM_1	72.2	65.8	32.3	24.5	62.5	52.3	59.4	51.6	47.8	39.7	53.2	45.9
	J_2	21.3	13.3	18.3	12.0	20.5	13.4	8.5	4.2	21.2	14.6	9.4	4.6
	LM_2	1.4	0.6	21.4	14.3	7.0	3.9	0.0	0.0	19.8	13.3	1.4	0.5
L=4:	J_1	78.5	72.5	71.1	64.1	82.9	77.3	51.6	46.4	73.0	66.0	57.8	49.6
	LM_1	61.6	49.9	52.2	35.7	34.0	19.2	43.0	37.9	74.7	69.1	42.3	31.5
	J_2	43.7	34.2	36.1	27.1	46.9	38.1	13.0	9.0	36.3	27.5	14.2	9.9
	LM_2	4.9	3.2	25.6	18.6	11.8	7.1	0.1	0.1	29.2	23.9	0.7	0.4
L=7:	J_1	83.3	77.6	75.6	69.5	87.6	82.2	71.2	64.6	82.2	76.3	90.1	85.1
	LM_1	45.8	30.0	15.4	7.2	20.4	12.3	44.2	37.8	77.2	72.0	45.1	30.4
	J_2	52.3	41.2	33.5	24.8	50.8	41.74	43.3	32.4	40.8	31.3	58.8	47.7
	LM_2	6.7	4.9	14.4	9.3	8.0	4.0	0.3	0.2	27.0	21.6	10.1	6.4
L=10:	J_1	89.7	86.1	84.8	79.7	93.4	89.9	83.5	77.1	90.5	87.1	95.2	92.3
	LM_1	45.2	31.9	11.3	5.7	21.8	15.5	41.1	35.3	79.2	73.4	35.0	23.9
	J_2	73.3	63.8	50.6	41.3	70.5	61.8	66.4	57.1	62.9	53.8	77.7	70.2
	LM_2	14.7	10.0	14.7	9.0	11.9	6.9	2.6	1.6	34.4	29.2	18.5	12.9
L=14:	J_1	92.8	89.3	88.7	83.5	95.9	93.2	86.2	80.8	92.1	89.5	95.5	93.0
	LM_1	35.4	23.7	10.6	5.5	20.7	14.2	36.7	31.8	78.4	71.0	21.0	13.2
	J_2	82.8	75.3	61.9	52.8	80.3	72.8	75.2	66.3	71.8	63.6	82.6	75.5
	LM_2	13.2	8.7	15.8	9.6	10.5	6.0	3.8	2.2	37.7	30.1	12.5	8.9

Note: *The MSE results for Ymodel 2 are multiplied by 100. Correct parametric specification for Ymodel 1 is φ_1, φ_2 for Ymodel 2 and φ_3 for Ymodel 3. Upper part of table provides for each model the MSE for the parametric estimator (L=0), and the MSE of the first GMM estimator (left column) and of the second GMM estimator (right column) for different numbers of nonparametric moments (L=1,4,7,10,14). The lower part of table contains the simulated size of the J and the LM tests at nominal size 10% (left column) and 5% (right column) for different numbers of nonparametric moments (L=1,4,7,10,14). J_1 and LM_1 are the J and the LM test based on the estimates of the first GMM estimator, J_1 =J($\hat{\theta}_1$), LM_1 =LM($\hat{\theta}_1$); J_2 and LM_2 are the J and the LM test based on the estimates of the second GMM estimator, J_2 =J($\hat{\theta}_2$), LM_2 =LM($\hat{\theta}_2$). 1000 replications.

The results are given in the lower part of Table 3.4 with J_1 and LM_1 referring to the tests

computed at the estimates $\hat{\theta}_1$ of the first GMM estimator and J_2 and LM_2 referring to the second GMM estimator. The column titled 10% provides the rejection frequency, using the 10%-critical value of the $\chi^2_{(L)}$ distribution with degrees of freedom equal to the number of subpopulations L. The column titled 5% contains accordingly the rejection frequency at the theoretical 5% level. It is seen that the J-tests strongly tend to over-reject in most cases, which may not be surprising in light of the previous findings in the literature. However, also the LM-tests depart considerably from their nominal size. At closest to its theoretical values comes the LM_2 test evaluated at the second GMM estimator with L=7 subpopulation restrictions, which still displays considerable under-rejection in Ymodel 1 and over-rejection in Ymodel 2. Hence, bootstrap versions of these tests should be used, as developed in Brown and Newey (2001) and Hall and Horowitz (1996). This is left for future research.

Regarding the power of these tests in case of misspecification it can be seen that both the J and the LM tests reject at a very high frequency. Particularly at sample size 2000 and few overidentifying moments L rejection rates are often close to 100%. (Results can be found in the supplementary appendix.)

4 Optimal choice among Swedish rehabilitation programmes

The developed GMM estimator is now applied to analyze optimal treatment choices among rehabilitation programmes in Sweden. The Swedish rehabilitation policy distinguishes between vocational and non-vocational rehabilitation and is directed towards individuals with reduced work capacity due to long-term sickness (of at least one month). Non-vocational rehabilitation contains medical rehabilitation as well as social rehabilitation for individuals with alcohol, drug or psychiatric problems and intends to re-establish independency of the sick individual from medical or therapeutical assistance. Vocational rehabilitation consists of workplace training and occupation-related educational measures and aims at restoring lost working capacity and re-integration into the labour market. A data set for the evaluation of the effects of vocational rehabilitation has been collected by the Swedish National Social Insurance Board, of which 6287 cases in Western Sweden are analyzed. The data set is very informative about the selection process, containing information on medical examination, medical recommendation, case workers' recommendations, the individual's sickness history and so forth, which renders the conditional independence assumption (3) plausible. For details on the data set, the selection of cases and the justification of the conditional independence assumption see Frölich, Heshmati, and Lechner (2000a). The various rehabilitation programmes are classified here as in Frölich, Heshmati, and Lechner (2000b) into 4 categories: *No rehabilitation*, *workplace rehabilitation*, *educational rehabilitation* and *medical&social rehabilitation*.

In a retrospective analysis for each of the 6287 sickness cases it is estimated which would have been the optimal programme choice. Comparing these estimates with the

observed participant allocation allows to assess by how much the process of allocating participants to programmes could be improved. The success of rehabilitation is measured by the *employment* status (employed/non-employed) at the end of the sickness spell, where employment could be either with a new or with the current employer. This measures the *short-term* reintegration into the labour market, which is one of the aims of vocational rehabilitation programmes. It should be noted that concentrating on a single outcome variable does not do justice to the multi-facetted goals of rehabilitation programmes, where for instance medical & social rehabilitation aims rather at improving health condition than rapid labour market reintegration. Also the durability of re-employment would be of interest. A more comprehensive analysis, however, was not possible due to the lack of follow-up data and identification concerns. In particular, conditional independence (3) seems not to be plausible with respect to after-treatment health status as pre-treatment health conditions are not reported as detailed as deemed necessary, see Frölich, Heshmati, and Lechner (2000a). Nevertheless, labour market reintegration is still of high interest from the viewpoint of an economist. (But policy conclusions should be cautiously drawn.)

Table 4.1 provides the number of participants in each rehabilitation group and the share of participants who engaged in employment at the end of their sickness spell. On average 46% of all cases are employed at the end of sickness. The average employment rate is 48% for the participants in No rehabilitation, 52% for the participants in workplace rehabilitation, about 29% for the participants in educational rehabilitation and 41% for the participants in medical rehabilitation.

Table 4.1: Treatment groups and their re-employment rate

	All	No Reha	Workplace	Educational	Medical
# Observations	6287	3502	1118	360	1307
Re-employment rate	46.3	48.3	52.4	28.9	40.5

Note: Share of transitions to employment at the end of sickness (in %).

Table 4.2: Nonparametrically estimated mean potential outcomes (in %)

Estimated	$\widehat{E[Y^{No}]}$	$\widehat{E[Y^{Work}]}$	$\widehat{E[Y^{Edu}]}$	$\widehat{E[Y^{Med}]}$
re-employment rate	46.0	45.6	32.9	41.0

Note: Mean potential outcomes estimated by SG propensity score matching (see Frölich (2000)).

However, the gross success rates of Table 4.1 are not particularly informative since the participants in the different programmes are rather different in their characteristics. Therefore, Table 4.2 provides the estimated mean potential outcomes $E[Y^{No}]$, $E[Y^{Work}]$, $E[Y^{Edu}]$ and $E[Y^{Med}]$, where the different compositions of characteristics are adjusted for by SG matching (Frölich 2000). In the absence of general equilibrium (macroeconomic) effects, these estimates can be interpreted such that, if all observations had participated in No or workplace

rehabilitation the employment rate would have been about 46%. On the other hand, had all observations instead participated in education rehabilitation 33% would have had encountered re-employment, and the re-employment rate would have been 41% if all observations had been allocated to medical rehabilitation instead. The differences between these estimates represent average treatment effects and indicate that educational rehabilitation fails completely in fostering re-integration in the labour market, as has also been observed in Frölich, Heshmati, and Lechner (2000b). This raises the question: Should educational rehabilitation be abolished completely? Or are there individuals for whom educational rehabilitation is the optimal programme and which would lose if this option were no longer available? And which employment rate could have been achieved if all individuals had been allocated optimally to the programmes?

To address these questions the optimal programme is estimated for each of the 6287 observations. With the outcome variable being binary the expectation function is specified as a probit

$$E[Y^r|X=x] \doteq \Phi(x'\theta^r) \qquad \forall r \in \{No, Work, Edu, Med\}, \tag{17}$$

where Φ is the cdf of the standard normal distribution and the coefficients θ^r may be different for each programme $r \in \{No, Work, Edu, Med\}$. The scores of the log likelihood function $\frac{\partial \ln l(x'\theta^r)}{\partial \theta^r} = \frac{\phi(x'\theta^r)}{\Phi(x'\theta^r)(1-\Phi(x'\theta^r))} x \cdot (y - \Phi(x'\theta^r))$ are taken as the instruments $A^r(X_i)$ in (12). The mean differences between parametric and nonparametric model according to 11 populations (see Table A.1 in appendix) are included in the GMM estimator. 38 explanatory characteristics (plus a constant) are included in x, comprising socioeconomic variables, indicators on sickness history and variables characterizing the current sickness spell. These variables were selected according to the analysis of the selection process in Frölich, Heshmati, and Lechner (2000b), augmented by variables that are considered relevant for predicting employment prospects. See Table A.2 in the appendix.[14]

The participation probabilities $\hat{p}^r$ are estimated by probit and the support restriction is implemented by discarding all observations with $\hat{p}_i^r$ below the lowest participation probability among the participants in programme r. The regression curves $m^r(p^r)$ are estimated for each subpopulation separately by SG matching, using only the observations belonging to that subpopulation. The bandwidth is chosen by least-squares cross validation. (The implied average potential outcomes for all subpopulations are given in Table A.1 in the appendix).

With these preliminary estimates the coefficients $\theta^{No}, \theta^{Work}, \theta^{Edu}, \theta^{Med}$ are estimated (one after the other) by the GMM estimator (13) using the identity matrix as weighting matrix,

[14] The same variables are also used for the mean potential outcomes in Table 4.2. Some variables that were included in Frölich, Heshmati, and Lechner (2000b) are left out here since they caused a singularity problem in the bootstrap simulation described below. These are: widowed, county Halland, county Göteborg (county Värmland included instead), indications of alcohol abuse, disability pension recommended by case worker, rehabilitation prevented by other factors, physician *and* case worker recommended a wait&see strategy.

as suggested by the Monte Carlo experiment.[15] With these coefficient estimates the expected potential outcomes $\hat{Y}_i^{No}, \hat{Y}_i^{Work}, \hat{Y}_i^{Edu}, \hat{Y}_i^{Med}$ are predicted for each observation and the optimal programme for observation i is the programme corresponding to the maximum of these four potential outcomes. Table 4.3 states for how many individual No, workplace, educational or medical rehabilitation is the estimated optimal programme. From this table it is seen that although No and workplace rehabilitation are optimal for the majority of all individuals, still for 1519 observations educational rehabilitation would have been the optimal choice. This might suggest that the complete elimination of educational rehabilitation might not be a wise choice, but that rather an improved allocation of participants might be recommendable.

Table 4.3: Distribution of optimal programme

Best programme is	No Reha	Workplace	Educational	Medical
for so many individuals:	1865	1860	1519	1043

Note: Number of individuals for whom No rehabilitation, workplace rehabilitation, educational rehabilitation or medical rehabilitation, respectively, is the estimated optimal programme.

Table 4.4: Distribution of optimal programme simulated by bootstrap

Best programme is	No Reha	Workplace	Educational	Medical	Undefined
with 90% probability for	142	100	23	16	6006 individuals
with 70% probability for	618	540	294	180	4655 individuals
with 60% probability for	920	893	552	352	3570 individuals
with 50% probability for	1302	1386	905	606	2088 individuals

Note: Number of individuals for whom the corresponding programme is the estimated optimal programme with probability 1-α=0.9, 0.7, 0.6, 0.5, respectively. 350 bootstrap replications.

However, in the analysis so far the sampling variability of the estimated coefficients has been neglected. To take this into account, the distribution of $\hat{\theta}^{No}, \hat{\theta}^{Work}, \hat{\theta}^{Edu}, \hat{\theta}^{Med}$ is simulated by bootstrap and a programme r is only defined as optimal for individual i if the simulated probability that programme r corresponds to the maximum of $\hat{Y}_i^{No}, \hat{Y}_i^{Work}, \hat{Y}_i^{Edu}, \hat{Y}_i^{Med}$ exceeds a certain threshold. According to (2) r_i^* is the optimal programme for individual i if

$$P\left(\arg\max_r \left\{\hat{Y}_i^{r=No}, \hat{Y}_i^{r=Work}, \hat{Y}_i^{r=Edu}, \hat{Y}_i^{r=Med}\right\} = r_i^*\right) \geq 1-\alpha.$$

The probability measure is simulated via bootstrap replications and a threshold of for instance 1-α=0.7 requires that in at least 70% of the bootstrap iterations r_i^* corresponds to the maximum of the estimated potential outcomes. Table 4.4 provides the number of individuals for whom No, workplace, educational or medical rehabilitation, respectively, is optimal with at least 90, 70, or 50% simulated probability. For all other individuals the optimal programme is undefined.

[15] The J and LM tests for θ^{No} are J_1: 45.2, LM_1: 55.6, J_2: 27.1 LM_2: 30.7, for θ^{Work} are J_1: 74.0, LM_1: 60.4, J_2: 12.3 LM_2: 23.9, for θ^{Edu} are J_1: 54.0, LM_1: 30.4, J_2: 21.7 LM_2: 23.0 and for θ^{Med} are J_1: 81.4, LM_1: 52.6, J_2: 27.4 LM_2: 40.3. Notation as in Table 3.4. According to the $\chi^2_{(11)}$ critical values all specifications are incorrect.

Table 4.4 shows again that for some individuals educational rehabilitation would be the optimal programme. Albeit a substantial amount of uncertainty is visible by the number of individuals without defined optimal choice even at the 0.5 level, optimal treatment choice seems to vary among the individuals and educational rehabilitation seems still to be the best choice for some individuals despite its weak average performance.

Table 4.5: Optimal treatment choice versus actual allocation

	r_i^* at 1-α=70%					r_i^* at 1-α=60%					r_i^* at 1-α=50%				
	N	W	E	M	i	N	W	E	M	i	N	W	E	M	i
D_i=N	399	198	156	113	2636	586	352	312	213	2039	828	603	507	351	1213
D_i=W	76	170	73	19	780	122	249	119	40	588	179	338	196	82	323
D_i=E	22	53	19	8	258	30	79	37	13	201	48	114	56	25	117
D_i=M	121	119	46	40	981	182	213	84	86	742	247	331	146	148	435
Total	618	540	294	180	4655	920	893	552	352	3570	1302	1386	905	606	2088
Δ(%)			61.5					64.7					67.4		

Note: Number of participants in programme $D_i \in \{No,Work,Edu,Med\}$ who have the same optimal programme $r_i^* \in \{No,Work,Edu,Med,indefinite\}$ at the level 1-α=0.7 (left), 0.6 (middle) and 0.5 (right). The column labelled i stands for indefinite optimal programme. Δ gives the fraction of misclassification in %, i.e. the number of cases for whom D_i and r_i^* do not coincide (off-diagonal elements) to the total number of cases with defined optimal programme, leaving apart the undefined cases.

It is revealing to compare this simulated optimal allocation with the actual allocation observed. Table 4.5 tabulates the number of participants in a certain programme $D_i \in \{No,Work,Edu,Med\}$ who have the same optimal programme $r_i^* \in \{No,Work,Edu,Med,indefinite\}$ at the 1-α level 70%, 60% and 50%, respectively. The row labelled Total gives the total number of individuals for whom a certain programme is determined as optimal. This row corresponds to the results of Table 4.4. For instance, for the 3502 individuals participating in No participation an optimal programme is determined in 866 cases at the 70% level, which is No rehabilitation in 399 cases, workplace rehabilitation in 198 cases, educational rehabilitation in 156 cases and medical rehabilitation in 113 cases. Particularly striking are the results for educational rehabilitation. Of the participants in educational rehabilitation only very few (19 cases) would have been assigned to educational rehabilitation under optimal allocation, whereas most of the 294 cases for which educational rehabilitation seems to be optimal actually participated in other programmes. These results are similar at the 60% and 50% level. This corroborates the finding that the participants in educational rehabilitation are not well selected. The fraction of misclassification Δ (in %) indicates how much optimal and actual classification deviate from each other. Leaving apart the cases for whom no optimal programme is defined, Δ is computed as the number of cases for which actual selection D_i and optimal choice r_i^* do not coincide (off-diagonal elements) divided by the total number of defined optimal programme choices. It is seen from Table 4.5 that at a probability level

of 0.7 more than 60% of the optimal programme choices differ from the actual allocation, indicating that substantial improvements might be possible by better programme selection. The misclassification level increases to 67% at the 0.5 level, which might be attributable to additional noise, since the optimal programme classifications are becoming less unambiguous.

Since the approach pursued in this paper estimates the optimal programme choice on an individual level it is difficult to summarize the optimal allocation by a few numbers or aggregate statistics. Nevertheless in Tables 4.6 and A.2 it is tried to spot some distinguishing trends between optimal and actual allocation. Table 4.6 provides the means of selected characteristics in the participant groups according to the actual and the optimal selection process, at 1-α level 50%. Table A.2 in the appendix shows this comparison for all 38 characteristics included in the estimator. In the first four columns the means among the different treatment groups according to the optimal allocation are given. (The 2088 individuals without defined optimal treatment are not included.) The column 'All' provides the means in the full sample and the last four columns provide the average characteristics among the actual participants. First, a striking difference with respect to age can be seen. Whereas the distribution of the age groups 18-35, 36-45 and 46-55 years among the actual participant groups does not depart very much from the distribution in the full sample, the optimal choice seems to depend strongly on the individual's age. Whereas the young are clearly over-represented among those who are advised to participate in medical and particularly in educational rehabilitation, only very few among the 46-55 years old are best served by educational rehabilitation. With respect to gender it seems as if men should more often attend No rehabilitation, whereas women might benefit more from workplace rehabilitation. Regarding prior unemployment it is noteworthy that hardly any unemployed are found among those advised to participate in No or in medical rehabilitation, but that they make up about half of those advised to educational rehabilitation. Educated blue collar workers are under-represented among the individuals for whom workplace rehabilitation appears to be optimal, whereas manufacturing workers are over-represented among those advised to No rehabilitation. For individuals who had been previously sick for more than 60 days in the last six months or who had participated in vocational rehabilitation medical rehabilitation is hardly ever an unambiguously optimal choice. Furthermore, in the optimal allocation individuals with psychiatric problems and individuals for whom a wait & see strategy has been advised are clearly under-represented in educational rehabilitation relative to the actual allocation observed. Generally the differences in the characteristics are much more pronounced in the optimal than in the actual allocation, indicating that the actual selection process is not very sensitive to observable characteristics

Table 4.6: Mean characteristics in treatment groups according to optimal and actual allocation

Variable		r_i^* =				All	D_i =			
		N	W	E	M		N	W	E	M
Age:	18-35 years	12	20	59	52	32	31	34	37	31
	46-55 years	40	62	10	30	37	41	31	32	36
Gender:	male	56	36	48	44	45	45	45	46	46
Employment status:	unemployed	2	27	47	2	19	20	9	32	21
Labour market position:	blue collar, high educated	43	9	19	17	20	20	23	23	20
Occupation in:	manufacturing	51	23	23	38	32	30	38	32	32
Previous sickness days	$>$ 60 days	19	32	25	5	22	20	24	35	22
Prior participation in	vocational rehabilitation	4	15	21	0	11	7	15	23	14
Medical diagnosis:	psychiatric	20	21	11	15	18	18	13	28	18
Medical recommend.	wait and see	79	64	19	53	55	61	40	37	56
Predicted employment	probability	69.3	52.6	54.5	67.2		48.5	51.9	30.2	41.1

Note: Average characteristics multiplied by 100. The first four columns refer to the optimal allocation (r_i^*) at level 1-α=0.5. The column labelled 'All' gives the mean for the full sample, and the last four columns provide the means among the actual participants (D_i). The last row presents the average predicted potential employment probabilities in the corresponding treatment groups.

In the last row of Table 4.6 the individually *predicted* potential employment outcomes are averaged within the participant groups according to the optimal and the actual allocation. The predicted average employment rates in the actual participant groups correspond quite well to the observed outcomes of Table 4.1. When re-allocating the participants to the programmes in an optimal way substantial increases in the predicted employment rates are achieved even for educational rehabilitation, as evidenced from Table 4.1. To summarize this analysis it is illuminating to predict tentatively the re-employment rate that would have been achieved had all individuals participated in their optimal programme. When allocating all individuals to their optimal programme, if defined at the 0.5 level, and all individuals without defined optimal programme randomly (with equal probability) to any programme the predicted average employment rate would be 54.5%. If instead the individuals without unambiguous optimal programme are allocated randomly to either No or workplace rehabilitation the predicted employment rate would be 55.7%. Thus, compared to the current selection process and to the re-employment rates that would be expected if all individuals were assigned to the same programme (see Table 4.2) an increase in the re-employment rate of about 9%-points seems to be possible through improved participants selection.

This also allows to re-assess the question whether educational rehabilitation should be eliminated completely. If educational rehabilitation were no longer available the predicted average employment rate would be 54.9%, when the individuals without unambiguous optimal programme are again assigned randomly to either No or workplace rehabilitation. Thus,

although educational rehabilitation is the optimal programme for quite many individuals, their second-best choice seems not to be much worse.

Similar results are also obtained for different variable and moment specifications. Comparing the above derived optimal allocation (with 11 population moments) with the optimal allocations that would result if 1, 6, 16 or 21, respectively, population moments were included, it is seen that the fraction of misclassification Δ (in %) between the main specification and any of these other specifications is at most 0.1% at the 1-α=0.7 level, at most 2.4% at the 0.6 level and at most 11% at the 0.5 level. Also compared to the parametric estimator (with 0 overidentifying moments) the number of misclassifications is low (1.1%, 4.1% and 14.1% at the 0.7, 0.6, 0.5 level, respectively). If the set of 11 population moments is maintained but the set of explanatory variables affected the estimated optimal allocation change more markedly. With a set of 28 or 30 variables the resulting allocations are still very similar (Δ of about 0.5%, about 5% and about 14.5% at the 0.7, 0.6, 0.5 level, respectively), but when leaving out relevant information on sickness history, diagnosis, geographic location, etc. (and retaining only 24 variables), the misclassification rates increase to 15.8%, 26.4% and almost 40%, respectively. Hence, detailed information might be necessary to obtain informed programme choices. (All tables are found in the supplementary appendix.)

In light of these, rather robust, results it seems as if educational rehabilitation could be abolished without any harm and that more care should be dedicated to participant selection and compliance.

5 Conclusions

In this paper a new semiparametric approach to estimating optimal programme choices is developed, which is based on combining a parametric specification for the conditional mean function with nonparametric estimates of average treatment outcomes. The results of a Monte Carlo experiment indicate that precision gains in estimating conditional potential outcomes can be achieved by including the nonparametric estimates into the estimator. With this estimator the optimal treatment choice among Swedish rehabilitation programmes with respect to the re-integration of long-term sick into the labour market is analyzed. It is found that the optimal programme is not the same for all individuals, and that optimal choices and observed choices often do not coincide. Even educational rehabilitation, despite a large negative treatment effect relative to all other programmes (including 'no rehabilitation'), is estimated to be the optimal programme choice for quite many individuals. When all individuals are allocated to their optimal treatment the re-employment rate is estimated to be about 56%, up from the 46% re-employment rate observed. This increase in the employment rate could not have been achieved by allocating all individuals to the same treatment, e.g. to workplace or to no-rehabilitation. Hence a diversity of different treatment options is valuable, but the

potential benefits due to targeted participant selection are thus far neglected. On the other hand, the potential benefits of educational rehabilitation seem to be very small. If educational rehabilitation were no longer available the estimated maximal employment rate drops only little to about 55%.

As a general conclusion it seems that the analysis of treatment effect heterogeneity is important, particularly for forward-looking policy recommendations. Allocating the right persons to the right programmes might be as important as the programmes themselves. Thus a general ranking of programmes into performing and non-performing might be a poor guidance for policy improvements. Using the approach developed in this paper it can be simulated which overall outcomes might be achievable by changing the treatment selection process and by how much the optimal outcomes might change if certain programme options are eliminated. In an ex-post consideration this approach can be used to assess the efficiency of the selection process by the mismatch between actual and optimal treatment choice, which could be utilized as a performance indicator for monitoring and counselling agencies, or to assess the relative merits of self-selection versus assignment.

However, this approach might also be useful in an ex-ante sense as a decision support system for decentralized decision makers, e.g. case workers choosing among different treatments, unemployed choosing between different training programmes, school-leavers choosing a profession. Providing them with detailed information about their expected potential outcomes (e.g. via the internet) might improve their programme choices, particularly if estimates on a variety of different outcome variables are delivered.

A Appendix: Swedish Rehabilitation Programmes

This appendix contains additional tables regarding the estimation of optimal programme choices. Further information about the data can be found in Frölich, Heshmati, and Lechner (2000a). Additional results of alternative specifications are available in the supplementary appendix on www.markusfroelich.de. In Table A.1 the observed treatment outcomes and the nonparametrically estimated counterfactual potential outcomes for the non-participants are displayed for the 11 populations used in the GMM estimator in Section 4. The entry in the second row/third column indicates that among the 3502 participants in No rehabilitation a employment rate of 48.3% was observed, while it was 49.8% among the participants in No rehabilitation aged 46-55 years. In the fourth column the expected potential outcome for No rehabilitation for those who had not participated in No rehabilitation is given for the different subpopulations. For instance, the counterfactual employment rate for No rehabilitation is 41.5% for the 46-55 years old who participated in workplace, educational or medical rehabilitation. These mean counterfactual outcomes are estimated separately for each population by SG matching (Frölich 2000) with the bandwidth value chosen by least-squares

cross-validation from the bandwidth grid $\{0.02, 0.04, .., 1\}$.

These estimates provide some indications about heterogeneous response among individuals: Whereas the re-employment rate according to No rehabilitation is as high as 52.6% among those participants who had previously been sick for less than 15 days, it is estimated to be only 32.5% among those occupied in agriculture who did not participate in No rehabilitation. Regarding educational rehabilitation, only 20.2% of the participants with psychiatric problems become immediately re-employed, whereas a counterfactual re-employment rate of above 38% is predicted for the non-participants who had previously been sick for less than 15 days or who live in the Älvsborg. Generally, the participants in No and in workplace rehabilitation seem to enjoy better re-employment chances than the participants in educational or medical rehabilitation irrespective of the treatment received, since their observed outcomes are in all subpopulations higher than the corresponding counterfactual outcome for the non-participants. For medical rehabilitation, on the other hand, this relationship is in most cases inverse. Despite these indications of heterogeneity, educational rehabilitation appears nevertheless in almost all cases as the worst programme.

Table A.1: Observed outcomes and estimated counterfactual outcomes for the 11 populations

		EY^N	$\hat{E}Y^N$	EY^W	$\hat{E}Y^W$	EY^E	$\hat{E}Y^E$	EY^M	$\hat{E}Y^M$
(Sub)Population	obs	D=N	D≠N	D=W	D≠W	D=E	D≠E	D=M	D≠M
All	6287	48.3	43.6	52.4	44.2	28.9	33.1	40.5	41.2
Age 46-55 years	2354	49.8	41.5	56.0	50.8	21.7	20.9	38.6	39.3
Occu in agriculture	1921	38.5	32.5	50.5	42.7	26.4	33.0	30.9	29.5
Previous sickness < 15d	3725	52.6	47.0	57.4	47.2	34.7	38.3	47.1	47.3
Previous sickness > 60d	1374	36.3	35.9	44.6	39.6	23.8	27.8	27.0	26.1
No previous VR participation	5611	49.7	45.3	53.3	44.8	29.6	32.3	42.7	42.9
County: Älvsborg	1829	47.2	39.7	56.5	49.5	32.8	38.6	31.8	35.7
County: Värmland	1470	46.1	44.7	49.5	41.5	28.3	28.2	39.8	43.9
Sickness in 1992/93	2203	48.6	42.6	54.6	47.7	30.8	28.8	41.1	40.5
Diagnosis: psychiatric	1102	41.6	36.1	47.6	38.7	20.2	21.3	30.6	33.5
Sickness registered by health care centre/hospital	5041	50.5	46.3	54.6	47.1	29.4	35.0	42.6	42.0

Note: obs = Number of observations in each subpopulation; $E[Y^{No}|D = No]$, $E[Y^{Work}|D = Work]$, $E[Y^{Edu}|D = Edu]$, $E[Y^{Med}|D = Med]$ are the average observed employment rates among the respective participants; $\hat{E}[Y^{No}|D \neq No]$, $\hat{E}[Y^{Work}|D \neq Work]$, $\hat{E}[Y^{Edu}|D \neq Edu]$, $\hat{E}[Y^{Med}|D \neq Med]$ are the counterfactual employment rates among the respective non-participants, estimated by SG propensity score matching. VR means vocational rehabilitation (= workplace and educational rehabilitation). The bandwidth values selected by cross-validation for the estimation of the Y^{No} potential outcome for the various subpopulations are: 0.16, 0.20, 0.16, 0.10, 0.58, 0.60, 0.16, 0.16, 0.38, 0.14, 1.00, respectively. For the estimation of Y^{Work} the bandwidths are: 0.06, 1.00, 1.00, 0.14, 1.00, 0.64, 0.06, 0.06, 1.00, 0.06, 1.00; for Y^{Edu}: 1.00, 0.14, 0.22, 0.14, 1.00, 1.00, 0.80, 0.50, 0.12, 1.00, 0.44; and for Y^{Med}: 0.62, 0.50, 0.06, 0.10, 0.74, 0.04, 1.00, 0.58, 0.78, 0.46, 0.66.

Table A.2: Average characteristics in treatment groups according to optimal and actual allocation

Variable		r_i^* =				All	D_i =			
		N	W	E	M		N	W	E	M
Age:	18-35 years	12	20	59	52	32	31	34	37	31
	46-55 years	40	62	10	30	37	41	31	32	36
Gender:	male	56	36	48	44	45	45	45	46	46
Citizenship:	Swedish born	85	88	83	87	86	86	88	90	83
Employment status:	unemployed	2	27	47	2	19	20	9	32	21
Income	(in SEK/1000)	1.4	1.2	1.4	1.1	1.3	1.3	1.3	1.3	1.3
Labour market position:	blue collar, low educated	36	57	37	50	45	42	52	47	47
	blue collar, high educated	43	9	19	17	20	20	23	23	20
	white collar	18	23	24	26	23	26	20	16	21
Occupation in:	health care	5	7	20	8	10	9	11	10	11
	various sciences	27	30	12	38	28	30	25	25	25
	manufacturing	51	23	23	38	32	30	38	32	32
Previous sickness days	31-60 days	15	11	0	7	10	9	9	10	11
(in last 6 months):	$>$ 60 days	19	32	25	5	22	20	24	35	22
Prior participation in	vocational rehabilitation	4	15	21	0	11	7	15	23	14
County:	Bohuslän	32	21	27	19	25	27	17	24	30
	Älvsborgslän	26	38	42	8	29	32	42	32	10
	Värmlandslän	21	17	22	41	23	23	29	29	18
Community type:	urban / suburban region	37	23	23	13	26	31	17	21	21
	major / middle large city	13	15	12	10	14	13	11	11	21
	industrial city	9	18	14	5	12	10	14	11	16
Unemployment rate	(in %)	6.4	7.0	6.2	6.4	6.5	6.5	6.6	6.7	6.6
Sickness registration by	psych./social med. centre	9	5	5	14	8	7	6	14	10
	private or other	6	9	15	21	12	11	13	13	11
Sickness degree:	100% sick leave	94	92	88	62	86	84	92	91	86
Medical diagnosis:	psychiatric	20	21	11	15	18	18	13	28	18
	musculoskeletal	40	46	45	48	44	39	51	44	51
	injuries	28	6	18	7	14	15	15	11	12
	other	4	22	19	16	15	18	13	10	12

continued on next page

Variable		r_i^* =				All	D_i =			
		N	W	E	M		N	W	E	M
Case assessed by:	the employer	28	30	15	13	23	17	40	25	25
	insurance office	14	23	23	4	16	13	16	33	22
	IO on behalf of employer	11	9	6	23	11	8	14	13	17
	not needed	21	13	28	49	26	36	10	9	16
Medical	wait and see	79	64	19	53	55	61	40	37	56
recommendation	VR needed and defined	10	27	42	32	26	14	47	55	34
Case worker recomm.:	VR needed and defined	11	51	39	35	32	17	63	62	38
Medical reasons	prevented VR	35	20	27	15	25	23	22	23	32
Med. & case worker rec.	VR needed and defined	10	15	29	25	19	9	35	44	25

Note: Average characteristics multiplied by 100 (except income). The first four columns refer to the optimal allocation (r_i^*) at level $1-\alpha = 0.5$. The column labelled 'All' gives the mean for the full sample, and the last four columns provide the means among the actual participants (D_i). VR stands for vocation rehabilitation, rec. means recommendation.

References

Altonji, J., and L. Segal (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14, 353–366.

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

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Angrist, J., and A. Krueger (1999): "Empirical Strategies in Labor Economics," in *The Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1277–1366. North-Holland, New York.

Back, K., and D. Brown (1993): "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–976.

Berger, M., D. Black, and J. Smith (2001): "Evaluating Profiling as a Means of Allocating Government Services," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 59–84. Physica/Springer, Heidelberg.

Björkland, A., and R. Moffitt (1987): "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69, 42–49.

Black, D., J. Smith, M. Berger, and B. Noel (1999): "Is the Threat of Training more Effective than Training itself?," *University of Western Ontario, Department of Economics Working Papers*, 9913.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32, 549–576.

Brodaty, T., B. Crépon, and D. Fougère (2001): "Using matching estimators to evaluate alternative youth employment programmes: Evidence from France, 1986-1988," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 85–124. Physica/Springer, Heidelberg.

Brown, B., and W. Newey (2001): "GMM, Efficient Bootstrapping, and Improved Inference," unpublished manuscript, Rice University and MIT.

Brown, B., W. Newey, and S. May (2001): "Bootstrapping with Moment Restrictions," unpublished manuscript, Rice University and MIT.

Burnside, C., and M. Eichenbaum (1996): "Small Sample Properties of Generalized Method of Moments based Wald Tests," *Journal of Business and Economic Statistics*, 14, 294–308.

Colpitts, T. (1999): "Targeting Reemployment Services in Canada: The Service and Outcome Measurement System (SOMS) Experience," mimeo, Department of Human Resources Development, Ottawa, Canada.

de Koning, J. (1999): "The chance-meter: Measuring the Individual Chance of Long-term Unemployment," Netherlands Economic Institute, Rotterdam.

Dehejia, R. (1999): "Program Evaluation as a Decision Problem," *NBER working paper*.

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

DOL (1999): *Evaluation of Worker Profiling and Reemployment Services Policy Workgroup: Final Report and Recommendations*. U.S. Department of Labor, Employment and Training Administration, Washington D.C.

Eberts, R. (1998): "The Use of Profiling to Target Services in State Welfare-to-Work Programs: An Example of Process and Implementation," *W.E. Upjohn Institute for Employment Reserach Working Paper*, 98-52.

Eberts, R., and C. O'Leary (1999): "A Frontline Decision Support System for One-Stop Career Centers," mimeo, W.E. Upjohn Institute for Employment Reserach.

Fan, J. (1992): "Design-adaptive Nonparametric Regression," *Journal of American Statistical Association*, 87, 998–1004.

Frölich, M. (2000): "Nonparametric Covariate Adjustment: Pair-matching versus Local Polynomial Matching," *Discussion Paper, Department of Economics, Universität St. Gallen*, 2000-17.

Frölich, M., A. Heshmati, and M. Lechner (2000a): "A Microeconometric Evaluation of Rehabilitation of Long-term Sickness in Sweden," *Discussion Paper, Department of Economics, Universität St. Gallen*, 2000-04.

——— (2000b): "Mikroökonometrische Evaluierung berufsbezogener Rehabilitation in Schweden," *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 136, 433–461.

Gerfin, M., and M. Lechner (2000): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Discussion Paper, Department of Economics, Universität St. Gallen*, 2000-10.

HALL, P., AND J. HOROWITZ (1996): "Bootstrap Critical Values for Tests based on Generalized-Method-of-Moments Estimators," *Econometrica*, 64, 891–916.

HANSEN, L. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

HÄRDLE, W. (1991): *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

HASTIE, T., AND C. LOADER (1992): "Local Regression. Automatic Kernel Carpentry," *Statistical Science*, 8, 120–143.

HECKMAN, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings*, 80, 313–318.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HECKMAN, J., R. LALONDE, AND J. SMITH (1999): "The Economics and Econometrics of Active Labour Market Programs," in *The Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. North-Holland, New York.

HECKMAN, J., AND R. ROBB (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labour Market Data*, ed. by J. Heckman, and B. Singer. Cambridge University Press, Cambridge.

HECKMAN, J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

HECKMAN, J., AND E. VYTLACIL (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings National Academic Sciences USA, Economic Sciences*, 96, 4730–4734.

HORRACE, W., AND P. SCHMIDT (2000): "Multiple Comparisons with the Best, with Economic Applications," *Journal of Applied Econometrics*, 15, 1–26.

HSU, J. (1996): *Multiple Comparisons: Theory and Methods*, vol. 1. Chapman and Hall, London.

IMBENS, G. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

IMBENS, G., AND J. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G., R. SPADY, AND P. JOHNSON (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333–357.

INKMANN, J. (2001): *Conditional Moment Estimation of Nonlinear Equation Systems*. Springer Verlag, Berlin.

JALAN, J., AND M. RAVALLION (2002): "Estimating the Benefit Incidence of an Antipoverty Program by Propensity Score Matching," *Journal of Business and Economic Statistics*, fortcoming.

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

——— (2000): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany," *Journal of Human Resources*, 35, 347–375.

——— (2001): "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 43–58. Physica/Springer, Heidelberg.

MANSKI, C. (1993): "The Selection Problem in Econometrics and Statistics," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science Publishers.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

——— (2000a): "Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415–442.

——— (2000b): "Using Studies of Treatment Response to Inform Treatment Choice in Heterogeneous Populations," *NBER, Technical Working Paper*, 263.

——— (2001): "Designing Programs for Heterogeneous Populations: The Value of Covariate Information," *American Economic Review, Papers and Proceedings*, 91, 103–106.

MOHR, L. (1999): "The Impact Profile Approach to Policy Merit," *Evaluation Review*, 23, 212–249.

OECD (1998): "The Early Identification of Jobseekers who are at Greatest Risk of Long-term Unemployment in Australia," in *Early Identification of Jobseekers at Risk of Long-term Unemployment: The Role of Profiling*, pp. 31–61. OECD Proceedings, Paris.

PUHANI, P. (1999): *Evaluating Active Labour Market Policies: Empirical Evidence for Poland during Transition*. Physica, Heidelberg.

ROSENBAUM, P., AND D. RUBIN (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.

SEIFERT, B., AND T. GASSER (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of American Statistical Association*, 91, 267–275.

——— (2000): "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, 9.

WALD, A. (1950): *Statistical Decision Functions*. Wiley, New York.