Evaluating Social Policy by Experimental and Nonexperimental Methods

by

Espen Bratberg*, Astrid Grasdal* and Alf Erling Risa*[§]

*University of Bergen, Dept. of Economics, Fosswinckelsgate 6,
N-5007 Bergen, Norway
*[§]Norwegian Centre in Organisation and Management, Rosenbergsgt. 39,
N-5015 Bergen, Norway

E-mail addresses: Espen.Bratberg@econ.uib.no, Astrid.Grasdal@econ.uib.no, Alf.Risa@econ.uib.no

Abstract: Establishing causal relationships in social policy evaluation is important, but difficult due to sample selection. To evaluate the performance of estimators designed to handle sample selection bias we analyse data from a Norwegian rehabilitation project with a randomised experimental design. The data permit us to compare the performance of different nonexperimental estimators with the experimental results. In our case study we find that nonexperimental evaluation based on sample selection estimators with selection terms which fails to meet conventional levels of statistical significance is highly unreliable. The difference in difference estimator and stratification on propensity scores perform better in our context.

The authors would like to thank Petter Bomann-Larsen, Ellen Haldorsen, Karsten Kronholm, Jan Sture Skouen, Holger Ursin, Jan Utkilen, and May Vik for cooperation in the Bergen experiment. Kjell Vaage and Arild Aakvik have contributed advice and discussions. The paper has clearly benefited from several useful comments from two referees and the editor. Risa would also like to thank the CES centre, Munich, for excellent working conditions during a stay in spring 1999. Financing by the Ministry of Social Affairs and the Norwegian Research Council is gratefully acknowledged.

I. Introduction

The problem of establishing firm causal relationships is a fundamental difficulty in social policy evaluation. To perfectly observe the effect of being exposed to a particular social policy, the investigator would like to know what the outcome for the individual is with the programme as compared with the outcome the same individual would have had without the programme. It is impossible to observe the same individual in both states, so auxiliary methods are needed. The standard textbook way to deal with such problems of causality is to perform a randomised controlled experiment. Randomised assignment in principle secures that observed and unobserved characteristics of the individuals in the treatment and control groups are equal. The only difference between the two groups is that one receives the treatment. Differences in average outcomes after treatment are therefore caused by the treatment, while reverse causality is ruled out.

For instance, a crude application of this method might imply randomly selecting a fraction of those seeking social assistance, and give them nothing. Monitoring the receivers and the non-receivers over time would reveal short and long term outcome effects of the programme. In most cases ethical and other considerations preclude this research strategy, and social policy researchers must rely on nonexperimental sample selection methods to address such issues. Therefore, sample selection methods are widely used for evaluation purposes in economics and social science, although it is not clear how well these methods recover causal relationships in practice. On the other hand, not only ethical objections have been raised against experimental designs for social policy evaluation either. Heckman and Smith (1995) point to several other problems of a methodological nature that may plague randomised experiments in social policy evaluation; while Burtless (1995) maintains that only randomised assignments guarantee that observed correlation between treatments and outcomes really reflect causal relationships. He also refers to the widely cited studies of LaLonde (1986) and

Fraker and Maynard (1987). They used the randomised trials in the American National Supported Work Demonstration, and compared actual estimates obtained in the demonstration with nonexperimental estimates that could be obtained by standard econometric methods if the control group in the demonstration were not available. Both these analyses concluded that the statistical methods of controlling for sample selection bias were not able to replicate the experimental effects.

Heckman and several co-authors have countered the criticism in several ways. One set of questions concerns the application and development of nonexperimental techniques. Heckman and Hotz (1989) assert that LaLonde (1986) and Fraker and Maynard (1987) did not apply available specification test for their nonexperimental studies. They show that such an exercise would have improved the results. Heckman et al. (1998) present new and improved methods of nonexperimental evaluation that rely less on untestable parametric model specifications. In later years nonexperimental evaluation methods have put a greater emphasis on sampling and matching procedures. This development is also evident in the papers evaluating evaluation methods by Friedlander and Robins (1995) and Dehejia and Wahba (1999).

In another set of questions, Heckman and Smith (1995) list several problems with randomised trials. Attrition from experimental samples implies that we observe outcome observations only for the self-selected sample that remain in the study. Therefore, selection issues that were the reason to have a controlled experiment to begin with may reappear in full strength. Another problem arises with randomisation bias. This occurs when the randomised assignment forces the individuals that participate in the programme to have different characteristics as compared with normal operating conditions. The last problem with experiments we mention on our short list of Heckman and Smith’s much longer one, is the case of substitution bias. This occurs when members of the control group gain access to close substitutes for the experimental treatment. All these arguments are relevant for a discussion over the relative merits of experimental vs. nonexperimental evaluation methods.

We relate to this literature by analysing data from a Norwegian randomised field trial aimed at vocational rehabilitation. The experiment randomised voluntary participants, but outcomes and background characteristics are also available for those that chose not to participate even though they satisfied the inclusion criteria of the experiment. We evaluate the performance of different evaluation methods in this setting by treating the non-participants as a comparison group. The objective of our study is to investigate whether nonexperimental estimators are able to recover the experimental average treatment effect of the treated. This case study may thus provide useful information for researchers that have to choose between nonexperimental research strategies when experimental data are not available.

The rest of the paper is organised in a conventional way. The next section contains a description of the Bergen experiment and the data. Then we turn to some econometric issues. Section four contains the empirical analyses where experimental results are compared to different nonexperimental model specifications, while concluding remarks sum up our findings.

II. The Bergen experiment

In the early 1990s, musculoskeletal pain accounted for approximately 45 % of Norwegian sickness spells lasting more than eight weeks, and for more than one third of all new entrants into disability pension every year. Workers on sick leave due to such complaints are normally followed up by a general practitioner (GP) and given some physical treatment, physiotherapy in particular. In the Bergen experiment, workers on sick leave due to musculoskeletal pain received treatment that, in addition to physiotherapy, also included a cognitive part aimed at increasing their knowledge about their health problems and increasing their ability and motivation to cope with them.[1] The main purpose of the experiment was to investigate if such interventions improve the ability to uphold work. In order to identify treatment effects, the experiment was performed as a randomised controlled study.

Experimental Design

The experiment included workers on sick leave for eight weeks or more with diagnoses given by a GP indicating back pain, neck/shoulder pain, general muscle pain and other conditions of more localised musculoskeletal disorders. In addition to the medical criterion, inclusion required that participants held a permanent job (full time or part time). Participants were recruited from the approximately 285 000 persons living in Bergen or in one of the five surrounding municipalities. During the enrolment period from November 1993 to March 1995 those who met the inclusion criteria were contacted in writing by the local social insurance authority inviting them to participate in the experiment. In the invitation letter it was emphasised that participation was voluntarily, that acceptance or rejection of the invitation would not affect sickness benefits, and that participation could mean assignment to a control group, which simply would imply ordinary treatment through their GP.

Workers who volunteered to receive the treatment first went through an examination performed by physiotherapists not involved in the treatment. This examination consisted of a set of standardised tests of functional ability and a medical/psychological questionnaire. Participants were then randomly assigned to treatment or to a control group.[2] Those assigned to treatment underwent a rehabilitation programme that lasted four weeks with six hours

[1] The type of treatment is documented in greater detail in Haldorsen et al. (1998).

[2] In order to ensure that the treatment groups always were filled, and that participants assigned to treatment never had to wait for more than one treatment period (five weeks), the allocation sequence followed an unequal randomisation of 2:1 in favour of the treatment group.

sessions five days a week. Treatment involved both individual and group interventions. In addition, participants in this group were followed up by the treatment team and given individual advice at three, six and ten months after they received treatment at the clinic. Participants in the control group were subjected to ordinary treatment by their GP without any systematic feedback or advice on therapy. After 12 months both the treatment and the control group underwent a new examination identical to the one performed before the random assignment.

Data

For those invited to the experiment (participants and non-participants) the National Insurance Administration (NIA) provides data from administrative records with information on timing and amounts of payments of sickness benefits, rehabilitation benefits and disability pension for a follow-up period of five years. Data from the NIA also include information on gender, date of birth/death, marital status, annual earnings, spouses annual earnings, and municipality of residence.

All Norwegian employees are covered by public sickness insurance. Workers on sick leave are entitled to sickness benefits for a period of maximum 12 months. Absences of more than three days must be certified by a physician. For a typical employee the replacement ratio is 100% from the first day of absence, with an upper replacement limit. The employer pays for the first two weeks.[3] Payments for absences exceeding two weeks are reimbursed by the NIA. After 12 months on sick leave one can apply for medical or vocational rehabilitation benefits or for disability pension. The rehabilitation benefit corresponds to a potential disability pension and compensates approximately 65 % of previous wage income. For administrative reasons sickness benefits for state employed civil servants are reimbursed at the institutional level. Hence, information regarding sickness spells is not available for this group of employees, which therefore had to be excluded from the analysis[4].

Of 1648 invited workers who met the inclusion criteria,[5] 560 accepted the invitation (participants), 498 responded negatively by returning an answer explaining that they did not want to receive the treatment (negative responders), and 590 did not respond at all to the invitation (non responders). In total, 358 participants were assigned to the treatment group and 202 to the control group. Of those assigned to treatment, 333 completed the program, 3 were excluded for medical reasons by the clinic while 22 withdrew from the programme before treatment was completed. The largest group of observations that were dropped due to missing data, lacked information on the diagnosis variable. This exclusion affects 4 participants and 66 non-participants. We expect that these exclusions will not affect the experimental estimates much, while improving the estimates of the health variables in the nonexperimental evaluations. To further improve the similarity of the experimental and the comparison groups, we constructed a common support for the estimated propensity score for receiving treatment among the treated and the comparisons.[6] In the main analysis we used those within the experimental and comparison groups whose estimated probabilities were in a common interval. Table 1 gives a brief overview.

(Table 1 about here)

The enrolment period lasted for several months in this experiment. We use the second calendar month after the month of invitation as a common starting point for the evaluation of

[3] The period of employer compensation has recently been increased to 16 days. This change does not affect our observation period.

[4] This problem is exclusively related to workers in the public sector who are employed by the central government. Workers in public sector employed at the municipal or county level are registered with individual records of sickness benefits.

[5] One individual declined having any information regarding him/herself collected for evaluation purposes and is treated as not meeting the inclusion criteria.

[6] The probit estimates yielding the propensity scores are reported in the Appendix.

both participants and non-participants.[7] For all subsequent calendar months we register whether the individual receive some sick pay or related benefits. Currently, data from the NIA include December 1997, giving us complete follow-up periods of two years for all individuals.

(Table 2 about here.)

The summary statistics in Table 2 reveal that this experiment involves groups that are widely different from the ones analysed by LaLonde (1986). The NSW programme analysed by LaLonde (1986) was designed to help disadvantaged workers lacking basic job skills to move into the labour market. Average pre-programme incomes for the males in that sample was about $3000 in 1982 dollars. We have a majority of middle aged married females with pre-programme earnings of about $25000 measured in 1997 dollars. The average earnings of spouses, for those who are married, are somewhat higher. These are middle class workers that have experienced sickness spells due to back problems and pains. On average, they have a positive earnings trend prior to the program, so we do not have a pre-programme earnings dip in the sense of Ashenfelter (1978). This suggests that the individuals recruited to the project are not affected by long term problems translating into pre-program earnings losses. However, both the participants and the non-participants experience an income decline after being sick. The non-participants return to work at a higher rate as compared to participants, but there is no difference in the incidence of returning to work between the treated and the controls. There is a difference in the earnings difference, however, between the treated and the controls, such that the treated have an earnings reduction of NOK 11700 ($1300) more as compared to controls. This negative effect in terms of earnings, amounting to 6% of pre-programme

[7] According to the project administrators the pre-randomisation examination of participants took place approximately 4-8 weeks after invitations were sent out.

earnings, can be interpreted directly as an unbiased, unadjusted estimate of the programme effect. However, the effect fails to meet conventional levels of significance.

Outcome variables

To assess the impact of treatment we use both a continuous and a dichotomous measure of outcome. The dichotomous outcome variable indicates whether the individual has returned to work or not, and is based on sick leave status the eighteenth follow-up month. Workers who are registered with some sickness benefits, rehabilitation benefits or disability pension in parts of, or throughout the entire calendar month, are defined as not having returned to work. By assuming that individuals who do not receive sick pay or related benefits are back at work, we may, by mistake, include individuals who have withdrawn from the labour market among the returned workers. However, since all participants were employed at the time when they were enrolled in the experiment and therefore have a job to return to if they recover, and since they are entitled to benefits if they are sick, we do not consider this to be an important problem.

The continuous outcome variable is annual earnings two years after enrolment (earnings(+2)).

III. The evaluation problem

For expositional simplicity, we present the evaluation problem in a regression framework. [8] Consider the effect on a continuous outcome variable, Y, of being assigned to the treatment group. The simplest evaluation problem is to assess the effect on Y of some "treatment", denoted by a dummy variable T which equals 1 if the treatment is received, 0 otherwise. We assume that

[8] For a more general formulation, see e.g. Heckman et al. (1998).

(1) $$Y_i = \begin{cases} Y_{i0} = \mathbf{ß'}\,\mathbf{X}_i + e_{i0} \text{ if } T_i = 0 \\ Y_{i1} = \mathbf{ß'}\,\mathbf{X}_i + g + e_{i1} \text{ if } T_i = 1 \end{cases},$$

where b and **X** are vectors of parameters (including a constant term) and exogenous covariates. The model may be extended by letting T_i interact with $\mathbf{X}_i$ to allow for heterogenous programme effects. e is a random error term with zero expectation. This term picks up the effects of unobserved factors that may affect the outcome but are uncorrelated with the variables in **X**. The impact of treatment is $\Delta_i = Y_{i1} - Y_{i0}$. This impact is unobservable for a particular individual *i*, because *i* cannot be observed in both the treatment and the non-treatment state. Taking expectations, we obtain

(2) $E(\Delta_i) = E(Y_{i1}) - E(Y_{i0}) = g + B_i$, where

(3) $B_i = E(e_i \mid T_i = 1) - E(e_i \mid T_i = 0).$

We would like to separate the systematic effect of treatment, *g*, from the effect of unobserved factors, B_i. The problem is that T_i and e_i may be correlated. For instance, unobserved motivation may differ across those that receive the treatment and those that do not. Therefore the random term in general does not disappear upon taking expectations. B_i is inherently unobservable. To obtain an assessment of *g*, it is necessary to let the outcomes of non-treated individuals proxy the outcomes of treated individuals in the absence of treatment.

The typical evaluation problem is caused by the possibility that $B_i \neq 0$, leading to a bias when estimating the treatment effect. In a randomised experiment, this bias is zero by construction.[9] The fact that someone was randomly assigned to treatment should contain no

[9] As pointed out in the introduction, experimental evaluation of social programmes may raise other problems.

information about the error term. Hence, the treatment effect may be obtained as the difference in mean outcomes, or as the OLS estimate of g in the regression

(4) $\quad Y_i = \mathbf{\ss' X}_i + gT_i + e_i$.[10]

If randomisation is not available, the treatment effect must be evaluated by comparing treated individuals to a nonexperimental comparison group. In that case, one must expect that $B \neq 0$. The fact that an individual is assigned to treatment results from a selection: either the individual volunteered, or he was picked among applicants by a programme manager. In both cases, there is no reason to presuppose that $E(e \mid T = 1) = E(e \mid T = 0)$. Any effort to estimate the treatment effect without considering this problem may lead to biased results, unless $B = 0$ by chance.

A common approach to the selection problem is to assume that selection into the programme is governed by the latent regression

(5) $\quad T_i{*} = \mathbf{d'Z}_i + u_i; T_i = 1 \text{ if } T_i{*} > 0, 0 \text{ otherwise},$

where $\mathbf{Z}$ and d denote variables and coefficients. Assuming joint normality of e and u, equations (4) and (5) may be estimated by maximum likelihood or the Heckman (1979) two step estimator, which utilises the fact that with joint normality and var(u) normalised to 1,

(6) $$E(e \mid T) = \text{corr}(u,e)\sqrt{\text{var}(e)}\,l(\mathbf{d'Z}), \text{ where } l(\mathbf{d'Z}) = T\frac{f(\mathbf{d'Z})}{\Phi(\mathbf{d'Z})} + (1-T)\frac{-f(\mathbf{d'Z})}{1-\Phi(\mathbf{d'Z})},$$

[10] Even with random assignment these unbiased estimators may yield slightly different results in finite samples.

and f and Φ denote the density and cumulative density functions of the standard normal distribution. Hence consistent parameter estimates may be obtained from the regression

$$(7) \qquad Y_i = \mathbf{ß'X}_i + gT_i + b_l \, l(\hat{\mathbf{d}}'\mathbf{Z}_i) + v_i ,$$

where $\hat{\mathbf{d}}$ is obtained by probit estimation of eq. (5).

Two-step estimation under less restrictive conditions is possible. Several varieties of a semi-parametric procedure have been suggested, see Vella (1998) or Pagan & Ullah (1999) for recent overviews. They all build on replacing the assumption that the distribution of (e, u) is bivariate normal with

$$(8) \qquad E(e \mid \mathbf{Z}, T) = g(\mathbf{d}'\mathbf{Z}),$$

where $g(.)$ is an unknown function, and $\mathbf{Z}$ is the same vector of variables as in (5). To obtain a semi-parametric two-step estimator, $Pr(T = 1| \mathbf{Z})$ is estimated non-parametrically, i.e. using an estimator which does not rely upon a specific distribution for u. This yields an estimate of the index d**’Z**. In the second step the conditional expectation in (8) is used when estimating the equation of interest, (1). One approach is to approximate $g(.)$ by some series expansion, a suggestion that is particularly easy to implement is using powers of the estimated index (Newey, 1988). Another approximation is to weight the powers by the inverse Mill’s ratio (Newey et al., 1990)[11]. An alternative (Powell,1987; Robinson, 1988) is to estimate $E(Y_i \mid$ d**’Z**$_i$) and $E(\mathbf{X}_i \mid$ d**’Z**$_i$) nonparametrically by kernel regression and then differencing out the

[11] Heckman and Robb (1985) used a Fourier transform. Newey (1999) establishes conditions where including the index as a regressor in the outcome equation yields consistent estimates even if the index is estimated parametrically.

selection bias. For identification, these methods depend on at least one variable in **Z** not being present in **X**. Identification in the joint normality-based methods, on the other hand, does not rely on this exclusion restriction but is obtained through functional form. In practice, identifying variables improve the model.

When repeated observations are available, with at least one pre- and post-treatment observation for the treatment group, the treatment effect may be obtained by a first difference estimator. This estimator is obtained from the regression

$$(9) \qquad Y_{it} - Y_{i,t-1} = \text{ß}'(\mathbf{X}_{it} - \mathbf{X}_{i,t-1}) + g\,(T_{it} - T_{i,t-1}) + e_{it} - e_{i,t-1}.$$

For participants, T_{it} = 1 for post-treatment periods. An advantage with this estimator is that it is not necessary to estimate the selection rule (5), neither is it necessary to assume that B_i = 0. What is needed, is that

$$(10) \qquad E(e_{it} - e_{i,t-1} \mid T_{it} = 1) = E(e_{it} - e_{i,t-1} \mid T_{it} = 0),$$

which is equivalent to assuming that the growth rate in absence of treatment would be the same for the treated and the non-treated.

A different approach to estimating treatment impacts is to use propensity scores to control for selection on observed differences between the treatment and comparison groups (Rosenbaum & Rubin, 1983). The propensity score is just $Pr(T_i = 1 \mid \mathbf{Z}_i)$ estimated by some probability model. Based on those, one can split the sample into strata with scores in the same intervals, thus obtaining sub-groups of treated and comparisons that are similar on the characteristics that are mapped into the score. Alternatively, each of the treated individuals may be matched directly to the comparison with the closest value of the propensity score.

In the next section, the performance of the following nonexperimental estimators is evaluated against the experimental results that were obtained in the Bergen experiment: Unadjusted OLS, Heckman (1979) two-step (bivariate normal assumption), bivariate normal model estimated by maximum likelihood, semi-parametric estimator with series expansion using inverse Mill's ratios, semi-parametric estimator with powers of estimated index[12], difference in difference estimator, and a matching estimator based on stratification on propensity scores. We also consider a discrete outcome (work/no work) where a probit performed on the experiment is compared to a nonexperimental bivariate probit estimated by maximum likelihood. This estimator also relies on the assumption that the error terms of the outcome and selection equations are bivariate normal.

IV. Empirical results

As seen in Table 2, the programme on average had no effect on the re-employment rate, whereas there was a negative but insignificant effect on earnings. A discussion of what may be the substantial reasons why the programme effect was nothing near what was hoped for is postponed to the next section. We now present experimental and nonexperimental estimates of programme effects on post programme earnings(+2). The main results are given in tables 3 through 5.[13]

Table 3 contains experimental regression results in column (1). The experimental estimates of the average treatment effect for the treated are compared to nonexperimental estimators evaluating treatment effects where the comparison group consists of 415 negative

[12] The index $\mathrm{d}'\mathbf{z}$ was estimated by the Manski (1975) maximum score estimator.

[13] We have not corrected for the 5.7% of dropouts in the treatment group. Even though dropouts pose potentially large problems for experimental evaluations, we consider the fraction in our experiment to be small enough to

responders, and 455 non-responders. The nonexperimental sample in Table 3 excludes the 33 observations that were outside of a common support for the estimated propensity score for receiving treatment as explained in the discussion of Table 1.[14]

Table 4 reports results for the same estimators with a different model specification allowing for heterogenous treatment effects. Finally, Table 5 reports results from a matching estimator based on stratification on estimated propensity scores for the treated and untreated.

(Table 3 about here)

Bearing in mind that the outcome variable is scaled by 10^{-7}, we see that the estimated effect of treatment on post-treatment earnings differences is –9.000 (approximately $ 1000, with a standard error of 9000) in the OLS adjusted regression on the experimental sample. This compares well with the average observed earnings difference between treated and controls (-11.700) found in Table 2. The treatment effect in the nonexperimental OLS reported in column (2) is –15.000. This negative effect is fairly close to the experimental result, although more precisely estimated, indicating that sample selection on earnings differences does not play a prominent role here.

The OLS estimate is slightly biased downwards. To compensate for the downward bias, the parametric sample selection models should identify a small, negative correlation coefficient between the error terms of the treatment equation and the earnings difference equation, adjusting the estimate upwards. On the contrary, the estimated selection terms are all positive, adjusting the estimated treatment effects in the wrong direction. However, none of the estimated selection terms, including those found for the semiparametric model, are

leave the problem unaddressed. A scaling of the treatment effect by division by the fraction of non-dropouts would yield a correction factor of 1.06. Heckman, Smith and Taber (1998) treat the dropout problem in depth.

[14] The sample for the semiparametric estimator in column (6) is different since that estimator does not depend on the participation probit defining the propensity score. The sample for the difference in differences estimator is smaller due to missing observations on time dependent covariates.

statistically significant. This finding suggests estimating the programme effect without selection equations.

We have performed a number of tests to check the appropriateness of different estimators.[15] Heckman and Hotz (1989) and later Friedlander and Robins (1995) have argued that a good specification test is to investigate whether preprogramme outcomes can be explained by later treatment with the same models that are used to explain postprogramme outcomes. If the program participation dummy yields program effects which are significantly different from zero in models where the outcome variable is changed to preprogramme income, then the nonexperimental estimator fails the test, otherwise it passes. We are not able to reject any of the nonexperimental estimators reported in Table 3 with this test.[16]

We conclude that the sample selection models reported in Table 3 are not able to improve nonexperimental estimates since they all wrongly identify a positive selection effect and adjust the estimated treatment effect in the wrong direction. We note that among the two-step estimators, the semiparametric estimator reported in column (6) produce the closest estimate to the experimental result. The selection terms are, however, not significant even in this estimation. The difference in differences estimator reported in column (7) performs quite well with an estimated treatment effect of –15.000.

A nonexperimental evaluator that relied on the large but imprecisely estimated selection terms in the sample selection models would be led astray. However, a prudent evaluator, discarding results based on insignificant selection terms, would retreat to

[15] The normality assumption in the selection equation has been tested with a conditional moments test as suggested by Pagan and Vella (1989), also see Chesher and Irish (1987). The null hypothesis that the error term in the selection equation is normal was not rejected.

Parametric sample selection models may be identified by functional form. However, identification is improved by introducing identifying excluding restrictions. In our participation equation, the dummy variable Bergen plays this role. It has a positive ($p<0.08$) effect on participation, and an insignificant effect on the outcome.

[16] In the specification tests the outcome variable is Earnings(-1). To avoid collinearity, the earnings variable entering the participation equations is changed to Earnings(-2), and Earnings trend is changed to Earnings(-2) – Earnings(-3).

unadjusted OLS and the difference in differences estimator both producing similar results not far from the experimental baseline.

The model estimated in Table 3 builds on the assumption of a common treatment effect across individuals. To permit heterogenous treatment effects, we estimated a model with treatment interactions on the experimental sample. We found significant interactions on earnings variables, reported in Table 4. To save space we only present the coefficients relating to treatment effects, the selection terms, and the specification tests.

(Table 4 about here)

Treatment interactions on earnings variables are highly significant, and improve overall model fit. Now we run preprogramme specification tests where the estimator fails the test if any programme interaction produces significant effects on preprogramme earnings. All the nonexperimental interaction models are rejected at a 10 per cent critical level by the specification tests. Rejection of this model specification may be related to the high serial correlation for earnings at the individual level.

(Table 5 about here.)

Table 5 presents estimates based on stratification on propensity scores for programme participation among the treated and the comparisons. The treatment effects obtained either as weighted unadjusted differences between strata, or as regression adjusted differences, are satisfactory, and close to the unadjusted OLS results.

Notice that few propensity scores for treatment are over 0.5. Even though a much lower fraction than this actually received treatment, the distribution of the estimated propensity scores indicates that the participation probit is not doing a particularly good job in predicting programme participation. This conclusion is supported by the fact that the histograms of the estimated propensity scores of programme participation for the treated and the comparison group as depicted in Figure 1 are quite similar.

(Figure 1 about here)

(Table 6 about here)

Table 6 shows the estimates of effects on the job return probability.[17] Here the bivariate probit estimator is used to correct for sample selection. Results from the bivariate probit are compared to a probit estimate of the return to job probability on the experimental sample, and to a univariate probit estimate on the nonexperimental sample. While there is not a significant sample selection problem in the earnings differences, the back to work outcome is biased downwards in the unadjusted univariate probit as compared to the experimental estimate in Table 6. However, instead of identifying a negative rho to rectify the discrepancy, the estimated rho is large and positive. The estimate of the treatment effect is therefore adjusted in the wrong direction, grossly exaggerating the true effect. But a likelihood ratio test of H_0: rho=0 does not reject the null hypothesis. The test suggests re-estimating with an ordinary univariate probit, the results of which are found in the last column of the table. The point estimate of the treatment effect in the univariate probit is closer, but still far off the experimental point estimate. The poor performance of the bivariate probit may be related to identification problems if the Bergen dummy is a weak instrument to identify selection. However, in a linear probability model a Hausman test for exogeneity of treatment with Bergen as a dummy rejects the null hypothesis at the 10 per cent level (p-value = 0.098).

V. Conclusion

We believe that the Bergen experiment had a tightly controlled experimental design. However, the unsatisfactory treatment effects that were measured in the experiment may in part be due to general problems that potentially plague many experiments as discussed by

Heckman and Smith (1995). Randomisation bias may have played a role. The rehabilitation experts at the clinic complained that further exclusions from the treatment group based on professional judgement would have improved results. Substitution bias may also have been present. The fact that a randomised study was going on was well known among other local providers of rehabilitation services. The treatment in the experimental rehabilitation clinic is evaluated against "standard practice." It may be that those who delivered "standard practice" at the time considered the new clinic a competitor, and were spurred to do well not to appear as failures treating the comparisons of the new project.

However, our main aim here is to report how experimental and nonexperimental estimators assess the outcomes of the experiment, given this particular design of the experiment. Our design to compare nonexperimental estimators against experimental results differs from previous studies in important ways. Firstly, the data collection in the original experiment was designed explicitly to make comparison possible. Secondly, although the treated group has experienced health problems it does not consist of socially disadvantaged individuals. Thirdly, they all belong to the same local area and labour market. This accentuates the fact that we present a case study, and it may increase homogeneity between the self-selected participants as compared to the non-participating comparisons. Indeed, the predicted treatment effect on earnings in the unadjusted nonexperimental OLS in column (2) of Table 3 is not far away from the experimental result, suggesting that sample selection is not a great problem in our case. In such circumstances, an evaluator's most grave concern is to avoid type two errors, i.e. not to identify and adjust for sample selection effects when they are not present in an important way. In this perspective, the fact that none of the sample selection estimators identify significant selection terms is good news. It is also important to note that in

[17] We have also estimated the model with treatment interactions. They were insignificant, and were dropped.

the present case, nonexperimental evaluation based on sample selection estimators with insignificant selection terms, tuns out to be highly unreliable.

We argued in the introduction that the need for reliable nonexperimental evaluation is great. A knowledge-based ambitious social policy needs tools to assess policy outcomes even in situations where controlled experiments are infeasible. And even where theoretically feasible, randomised field trials will not always be chosen for economic and many other reasons. Some of the results in this case study are encouraging for the practice of nonexperimental evaluation. However, our findings altogether support the view of the received literature that nonexperimental methods should be used with caution. We also believe that our findings accentuate the need for further efforts into collecting data that permit evaluation of the evaluation methods.

References

Ashenfelter, O.: Estimating the Effects of Training Programs on Earnings. *Review of Economics and Statistics 60*, 47-57, 1978.

Burtless, G.: The Case for Randomized Field Trials in Economic and Policy Research. *Journal of Economic Perspectives 9*, 63-84, 1995.

Chesher, A. & Irish, M.: Residual Analysis in the Grouped and Censored Normal Linear Model. *Journal of Econometrics 34*, 33-61, 1987.

Deheija, R.H. & Wahba, S.: Causal Effects in Nonexperimental Studies: Reevaluation of the Evaluation of Training Programs. *Journal of the American Statistical Association, 94,* 1053-1062, 1999.

Fraker, T. & Maynard, R.: The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *Journal of Human Resources 22*, 194-227, 1987.

Friedlander, D. & Robins, P.K.: Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review 85*, 923-937, 1995.

Haldorsen, E.M.H., Kronholm, K., Skouen, J.S. & Ursin, H.: Multimodal Cognitive Behavioural Treatment of Patients Sicklisted for Musculoskeletal Pain: A Randomized Controlled Study. *Scandinavian Journal of Rheumatology 27*, 16-25, 1998.

Heckman, J. J.: Sample Selection Bias as a Specification Error. *Econometrica 47*, 153-161, 1979.

Heckman, J. J. & Hotz, V. J.: Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association 84*, 862-874, 1989.

Heckman, J. J. & Robb, R: Alternative Methods for Evaluating the Impact of Interventions. In J. Heckman & B. Singer (eds.), *Longitudinal Analysis of Labor Market Data,* Cambridge University Press, New York, 1985.

Heckman, J. J. & Smith, J. A.: Assessing the Case for Social Experiments. *Journal of Economic Perspectives 9*, 85-110, 1995.

Heckman, J. J., Ichimura, H., Smith, J. & Todd, P.: Characterizing Selection Bias Using Experimental Data. *Econometrica 66*, 1017-1098, 1998.

Heckman, J. J., Smith, J. & Taber, C.: Accounting for Dropouts in Evaluations of Social Programs. *Review of Economics and Statistics 80*, 1-14, 1998.

LaLonde, R. J.: Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review 76*, 604-620, 1986.

Manski, C.: The Maximum Score Estimator of the Stochastic Utility Model of Choice. *Journal of Econometrics, 3*, 205-228, 1975.

Newey, W.: Two- Step Series Estimation of Sample Selection Models. Princeton University, 1988.

Newey, W.: Consistency of Two-Step Sample Selection Estimator Despite Misspecification of Distribution. *Economics Letters*, *63*, 129-132, 1999.

Newey, W. K., Powell, J. L. & Walker, J. R.: Semiparametric Estimation of Selection Models: Some Empirical Results. *American Economic Review, Papers and Proceedings 80,* 324-328, 1990.

Pagan, A. & Vella, F.: Diagnostic Tests for Models Based on Individual Data: A Survey. *Journal of Applied Econometrics 4,* S29-S59, 1989.

Pagan, A. & Ullah, A: *Nonparametric Econometrics*. Cambridge University Press, Cambridge, 1999.

Powell, J.L.: Semiparametric Estimation of Bivariate Latent Variable Models. University of Wisconsin-Madison, 1987.

Robinson, P.: Root-n Consistent Semiparametric Regression. *Econometrica 56*, 931-954, 1988.

Rosenbaum, P. & Rubin, D.: The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70,* 41 – 55, 1983.

Vella, F.: Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources 33,* 127-169, 1998.

White, H.: A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica 48,* 817-838, 1980.

Table 1. *Sampling procedure*

	Total	Treated	Controls	Negative responders	Non responders
Invited Participants / Non participants	1648	358	202	498	590
Excluded :					
-exclusion at the rehabilitation clinic	3	3			
-state-employed civil servants	91	16	8	27	40
-low income prior to invitation	26	2	4	6	14
-dead or retired at age 67	21	6	3	5	7
-missing diagnosis	70	1	3	25	41
-missing data on other variables	42	12	6	9	15
Outside common support[a)]	33	3[b)]	1[b)]	11	18
Final Sample	1395	315(+3)	177(+1)	426	473
Withdrawals (in total)	22	22			
(in final sample)	18	18			

[a)] Constructed by probit estimation of treatment probability.

[b)] Not excluded in the experimental estimates.

Table 2 *Summary statistics for the experimental participants and non-participants.*

	Participants				Non-participants			
	Treated		Controls		Negative responders		Non responders	
	Mean	Std.dev	Mean	Std.dev	Mean	Std.dev	Mean	Std.dev
Male	0.37		0.37		0.35		0.50	
Age	43.5	(10.6)	43.3	(10.5)	46.0	(11.2)	42.3	(11.6)
< 30	0.10		0.13		0.09		0.14	
31-45	0.44		0.43		0.36		0.46	
46-55	0.27		0.29		0.31		0.23	
56-65	0.18		0.15		0.24		0.17	
Married	0.61		0.62		0.64		0.56	
Single	0.17		0.16		0.18		0.25	
Previously married	0.22		0.22		0.18		0.19	
Back pain	0.47		0.52		0.40		0.44	
Neck pain	0.15		0.16		0.15		0.16	
Muscle pain	0.11		0.07		0.09		0.07	
Other diagnosis	0.27		0.24		0.36		0.33	
Months on sick leave	3.2	(1.2)	3.1	(1.2)	2.9	(1.2)	2.9	(1.2)
Living in Bergen	0.87		0.85		0.85		0.79	
Earnings(-2) [b]	186.6	(79.0)	179.4	(84.1)	188.0	(89.3)	183.5	(96.3)
Earnings(-1) [b]	193.4	(76.4)	185.5	(77.71)	194.0	(102.1)	195.5	(89.1)
Earnings(0) [a]	189.2	(75.8)	183.1	(75.5)	192.5	(85.8)	205.5	(89.5)
Earnings(+2) [b]	151.8	(118.7)	155.6	(110.1)	167.4	(120.8)	170.0	(105.9)
Earnings trend	6.9	(36.4)	6.2	(44.3)	6.0	(48.0)	12.0	(68.3)
Spouse earnings (if married) [a]	198.0	(138.8)	195.5	(135.0)	204.7	(153.5)	209.3	(338.0)
% work18 [d]	49.4		50.6		61.3		60.3	
% work16-20 [d]	40.6		40.4		54.2		48.8	
Earnings difference [c]	-41.6	(96.0)	-29.9	(104.1)	-26.6	(92.2)	-25.5	(103.7)
# observations	318		178		426		473	

[a] Annual earnings in year of enrolment. All measures of earnings in this table are in NoK(1997)/10^3.

[b] -1 and -2 refer to annual earnings one and two years prior to enrolment year, whereas +2 refers to earnings the second year after enrolment year.

[c] Earnings(+2)-Earnings(-1)

[d] Sick leave status evaluated the 18th / 16th-20th calendar month after enrolment. Workers who do not receive sickness benefits, rehabilitation benefits or an increased disability pension (compared to an eventual pre-enrolment pension) during this/these month(s) are interpreted as returned to work.

Table 3. *Estimation of post-treatment income parameters (standard errors in pharentheses).*

Variables[a)]	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	-0.0009	-0.0015 **	-0.0089	-0.0107 ***	-0.0152	-0.0018 **	-0.0015 **
	(0.0009)	*(0.0006)*	*(0.0123)*	*(0.0022)*	*(0.0125)*	*(0.0009)*	*(0.0007)*
Age/100	-0.0163 ***	-0.0196 ***	-0.0210 ***	-0.0213 ***	-0.0234 ***	-0.0187 ***	
	(0.0047)	*(0.0028)*	*(0.0038)*	*(0.0031)*	*(0.0043)*	*(0.0029)*	
Male	-0.0008	0.0004	0.0001	0.0000	-0.0004	0.0007	
	(0.0011)	*(0.0006)*	*(0.0009)*	*(0.0007)*	*(0.0010)*	*(0.0006)*	
Months sick/10	-0.0049	-0.0059 ***	-0.0032	-0.0025	-0.0001	-0.0059 ***	
	(0.0035)	*(0.0023)*	*(0.0052)*	*(0.0026)*	*(0.0057)*	*(0.0022)*	
Backpain	-0.0004	-0.0003	0.0002	0.0003	0.0006	-0.0002	
	(0.0010)	*(0.0006)*	*(0.0010)*	*(0.0007)*	*(0.0011)*	*(0.0007)*	
Neckpain	-0.0013	-0.0022 ***	-0.0020 **	-0.0020 **	-0.0018 **	-0.0035 ***	
	(0.0014)	*(0.0008)*	*(0.0009)*	*(0.0009)*	*(0.0009)*	*(0.0013)*	
Generalised pain	0.0024	-0.0006	0.0001	0.0003	0.0009	-0.0003	
	(0.0016)	*(0.0010)*	*(0.0016)*	*(0.0011)*	*(0.0018)*	*(0.0010)*	
Married	-0.0007	0.0015 *	0.0019	0.0020 **	0.0025 *	0.0015	0.0023
	(0.0016)	*(0.0009)*	*(0.0012)*	*(0.0009)*	*(0.0014)*	*(0.0010)*	*(0.0017)*
Previously married	0.0003	0.0003	0.0010	0.0011	0.0018	-0.0020	-0.0011
	(0.0015)	*(0.0010)*	*(0.0015)*	*(0.0011)*	*(0.0017)*	*(0.0020)*	*(0.0015)*
Earnings trend	-0.0489	-0.1769 ***	-0.1917 ***	-0.1983 ***	-0.2122	-0.1698	
	(0.1111)	*(0.0551)*	*(0.0629)*	*(0.0598)*	*(0.1291)*	*(0.1294)*	
Earnings(-1)	0.1772	0.8773 ***	0.8842 ***	0.8769 ***	0.8972 ***	0.8729 ***	
	(0.1710)	*(0.0605)*	*(0.0626)*	*(0.0633)*	*(0.0815)*	*(0.0765)*	
Earnings(-1)^2	14.7520 ***	-1.9253 ***	-1.9130 ***	-1.7760 **	-1.8956	-1.9314	
	(3.3163)	*(0.7459)*	*(0.7444)*	*(0.7672)*	*(1.3275)*	*(1.3324)*	
Spouse earnings	0.1080 ***	0.0096	0.0097	0.0102	0.0099	0.0094	0.0417 **
	(0.0418)	*(0.0149)*	*(0.0149)*	*(0.0151)*	*(0.0149)*	*(0.0152)*	*(0.0188)*
Constant	0.0145 ***	0.0104 ***	0.0115 ***	0.0119 ***	0.0124 ***	0.0101 ***	-0.0025 ***
	(0.0028)	*(0.0016)*	*(0.0024)*	*(0.0017)*	*(0.0022)*	*(0.0018)*	*(0.0003)*
Selection terms							
Rho[b)]			0.4570	0.5503			
				(1.96)			
Lambda			0.0044		0.0101		
			(0.0073)		*(0.0078)*		
Index[c)]					0.0041	-0.0003	
					(0.0059)	*(0.0009)*	
Index^2[c)]					0.0012	0.0027	
					(0.0053)	*(0.0023)*	
R^2	0.3429	0.3685			0.3768	0.3782	0.010
Specification test[d)]	[0.859]	[0.353]	[0.691]	[0.420]	[0.962]	[0.614]	[0.208]
# observations	496	1185	1185	1185	1185	1202	1128

Note: Col.(1) Experiment, OLS; Col(2) Standard OLS; Col.(3) Heckman two-step; Col.(4) Sample selection MLE; Col.(5) Two-step series approximation weighted by inverse the Mill's ratio; Col.(6) Two step series approximation with the Manski maximum score index; (7) Difference in difference.

a) All earnings variables are in NOK(1997)*10^{-7}.

b) Estimated rho, Chi-square test statistic for LR test (rho=0) in parentheses.

c) Probit index weighted by lambda in column 5, maximum score index in column 6.

d) P-value of pre-treatment specification test.

Table 4. *Estimation of post-treatment income parameters in models with interactions (standard errors in pharentheses).*

Selected variables [a)]	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.0067	0.0064 **	0.0015	-0.0003	-0.0073	0.0057 **
	(0.0042)	*(0.0026)*	*(0.0122)*	*(0.0038)*	*(0.0125)*	*(0.0028)*
TxEarnings(-1)	-0.8392 **	-0.9413 ***	-0.9382 ***	-0.9516 ***	-0.9431 **	-0.9581 ***
	(0.3631)	*(0.2002)*	*(0.1970)*	*(0.1960)*	*(0.2599)*	*(0.2002)*
TxEarnings(-1)^2	22.5684 ***	22.4150 ***	22.3502 ***	22.3109 ***	22.5750 ***	22.6100 ***
	(7.6419)	*(3.7500)*	*(3.6854)*	*(3.6532)*	*(5.8380)*	*(3.7390)*
TxSpouse earnings	-0.0976	0.0478	0.0483	0.0496	0.0435 ***	0.0657
	(0.0606)	*(0.0387)*	*(0.0381)*	*(0.0378)*	*(0.0385)*	*(0.0419)*
Selection terms						
Rho [b)]			0.3115	0.4380		
				(0.88)		
Lambda			0.0029		0.0102	
			(0.0070)		*(0.0078)*	
Index [c)]					0.0035	-0.0008
					(0.0057)	*(0.0009)*
Index^2 [c)]					0.0003	0.0032
					(0.0050)	*(0.0023)*
R^2	0.3570	0.3888			0.3987	0.3914
Treatment effect [d)]	-0.0010	-0.0016	-0.0064	-0.0085	-0.0153	-0.0022
Specification test [e)]						
Treated	[0.383]	[0.032]	[0.805]	[0.255]	[0.653]	[0.058]
TxEarnings(-2)	[0.283]	[0.007]	[0.007]	[0.007]	[0.052]	[0.066]
TxEarnings(-2)^2	[0.117]	[0.006]	[0.006]	[0.006]	[0.110]	[0.139]
TxSpouse earnings	[0.244]	[0.570]	[0.572]	[0.570]	[0.609]	[0.365]
# observations	496	1185	1185	1185	1185	1202

Note: Col.(1) Experiment, OLS; Col(2) Standard OLS; Col.(3) Heckman two-step; Col.(4) Sample selection MLE; Col.(5) Two-step series approximation weighted by the inverse Mill's ratio; Col.(6) Two step series approximation with the Manski maximum score index..

*/**/*** indicates significance at 10%, 5% and 1% level.

a) Includes the same regressors as the model reported in table 3.

b) Estimated rho, Chi-square test statistic for LR test (rho=0) in parentheses.

c) Probit index weighted by lambda in column 5, maximum score index in column 6.

d) Mean of estimated effect of treatment on the treated.

e) P-value of pre-treatment specification test

Table 5. *Stratification on propensity score for programme participation, treated and comparison group members.*

				Mean, post treatment earnings			
						Weighted difference	
Strata [a)]	# obs.	# treated	Weight	Treated	Comparison group	Unadjusted	Regression adjusted
>0.15	36	5	0.016	0.01205	0.01685	-0.00008	-0.00005
0.15-0.20	218	40	0.127	0.01624	0.01738	-0.00014	-0.00014
0.20-0.25	308	71	0.225	0.01592	0.01697	-0.00023	-0.00010
0.25-0.30	275	79	0.251	0.01488	0.01695	-0.00052	-0.00098
0.30-0.35	198	64	0.203	0.01474	0.01670	-0.00040	-0.00011
0.35<	150	56	0.178	0.01447	0.01510	-0.00011	0.00014
# obs.	1185	315					
Treatment effect						-0.00149	-0.00124

a) Based on propensity scores obtained from probit regression of the probability of treatment, c.f. Appendix, table A2.

Table 6. *Estimation of parameters for post-treatment probability of having returned to work 18 months after enrolment, (estimated standard errors in parenthesis).*

	Experiment, Probit MLE		Bivariate probit MLE		Non-experimental probit MLE	
Variables	Coeff.	Std.err.	Coeff.	Std.err.	Coeff.	Std.err.
Treatment	-0.0319	*(0.1202)*	-1.3653	*(0.4287)* ***	-0.2650	*(0.0853)* ***
Age/100	-1.3411	*(0.6427)* **	-2.2726	*(0.4803)* ***	-2.5557	*(0.4093)* ***
Male	-0.0640	*(0.1464)*	-0.0225	*(0.0924)*	0.0407	*(0.0923)*
Months on sick leave/10	-1.2701	*(0.4740)* ***	-0.4862	*(0.4382)*	-1.0496	*(0.3233)* ***
Backpain	-0.1995	*(0.1421)*	-0.0521	*(0.0958)*	-0.1358	*(0.0896)*
Neckpain	-0.4829	*(0.1879)* ***	-0.2878	*(0.1259)* **	-0.3674	*(0.1174)*
Generalised pain	-0.2985	*(0.2159)*	-0.0240	*(0.1516)*	-0.1489	*(0.1465)*
Married	0.0573	*(0.2137)*	0.1850	*(0.1222)*	0.2254	*(0.1352)* *
Previously married	-0.0891	*(0.2041)*	0.1265	*(0.1181)*	0.1481	*(0.1354)*
Earnings trend	2.6279	*(15.2872)*	-14.0754	*(6.9757)* **	-13.8343	*(7.2076)* *
Earnings(-1)	23.0834	*(9.0366)* **	21.7477	*(5.2436)* ***	23.2234	*(5.1712)* ***
Spouse earnings	3.7052	*(5.6654)*	6.5945	*(3.0489)* **	7.4360	*(3.3674)* **
Constant	0.7225	*(0.3378)* **	1.1235	*(0.2184)* ***	1.1245	*(0.2057)* ***
Rho [a)]			0.689	(1.122)		
Treatment effect [b)]	-0.0109		-0.3116		-0.1084	
Log-likelihood	-329.050		-1429.1232		-753.919	
R^2	0.0429				0.0667	
# observations	496		1185		1185	

Note: Col.(1) Experiment, Probit MLE; Col(2) Non-experimental Probit MLE; Col.(3) Bivariate Probit MLE.

*/**/*** indicates significance at 10%, 5% and 1% level.

a) Estimated rho, Chi-squate test statistic for LR test (rho=0) in parentheses.

b)) =Pr(y_1=1)-Pr(y_0=1)

APPENDIX

Table A1 *Variable definitions*

Variable	Definition
Work18	Dummy indicating return to work 18 months after the pre-test
Work16-20	Dummy indicating return to work in month 16-20 after the pre-test
Sick	Number of months on sick leave during the last 6 months prior to enrolment
Age	Age when enrolled in the experiment
Bergen	Dummy indicating citizen of the municipality of Bergen
Married	Dummy indicating marriage
Previously married	Dummy indicating previously married (divorced, separated or widow(er))
Back pain	Dummy indicating back pain (ICPC diagnoses L02,L03,L84 or L86)
Neck pain	Dummy indicating neck/shoulder pain (ICPC diagnoses L01, L83, L83, L92)
Generalised pain	Dummy indicating general musculoskeletal pain (ICPC diagnoses L18, L19, L29, L99)
Other diagnosis	Dummy indicating other conditions of more localised musculoskeletal disorders (ICPC diagnoses L08, L09, L20, L81, L82, L85, L93, N02)
Earnings trend	Difference between earnings the year prior to enrolment year and earnings one year earlier
Earnings(i)	Annual earnings in year i = -2,-1,0,+1,+2 where 0 denotes year of enrolment in the experiment. All earnings measures are in NoK 1997*10^{-7} .
Earnings(i)^2	Earnings(i) squared
Spouse earnings	Spouses earnings in year of enrolment
TxEarnings(i)	Interaction between Treatment dummy and earnings
TxEarnings(i)^2	Interaction between Treatment dummy and earnings squared

Table A2 *Probability of treatment*.

	ML probit estimates		
Variable	Coef.	Std.err	dF/dz [a)]
Bergen	0.2062 *	*(0.1114)*	0.0630
Age /100	-0.8956 **	*(0.4172)*	-0.2875
Male	-0.1807 **	*(0.0919)*	-0.0574
Sick/100	1.2066 ***	*(0.3274)*	0.3873
Back pain	0.2283 **	*(0.0933)*	0.0739
Neck pain	0.0955	*(0.1240)*	0.0313
Muscle pain	0.3757 **	*(0.1475)*	0.1315
Married	0.2547 **	*(0.1193)*	0.0803
Previously married	0.3389 **	*(0.1399)*	0.1155
Earnings trend	-6.8118	*(7.7707)*	-2.1867
Earnings(-1)	4.2733	*(5.1335)*	1.3718
Constant	-1.1664 ***	*(0.2285)*	
lnL	-679.206		
# observations	1217		

a) Mean derivatives, finite differences for the dummy variables.

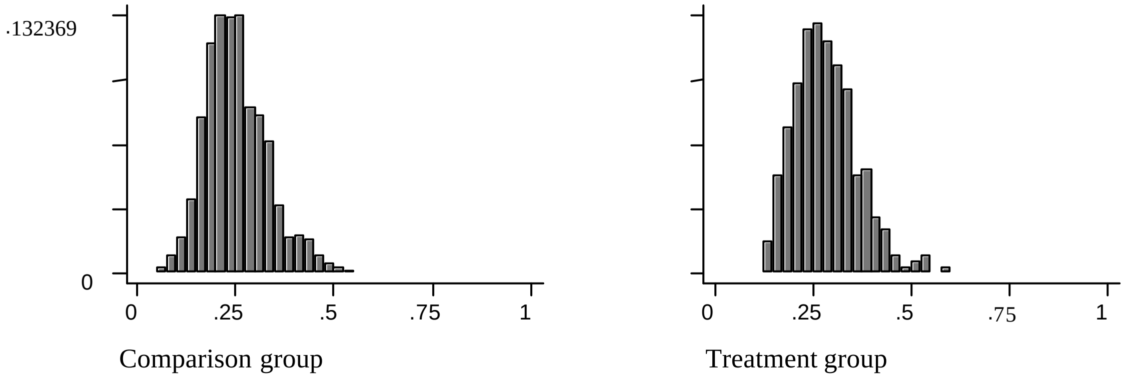

Fig.1 Histograms of predicted probability of treatment for treatment and comparison group members.