The Economic and Social Review, Vol. 31, No. 4, October, 2000, pp. 283-308

Evaluating State Programmes: "Natural Experiments" and Propensity Scores*

DENIS CONNIFFE
VANESSA GASH
PHILIP J. O'CONNELL
The Economic and Social Research Institute, Dublin

Abstract: Evaluations of programmes — for example, labour market interventions such as employment schemes and training courses — usually involve comparison of the performance of a treatment group (recipients of the programme) with a control group (non-recipients) as regards some response (gaining employment, for example). But the ideal of randomisation of individuals to groups is rarely possible in the social sciences and there may be substantial differences between groups in the distributions of individual characteristics that can affect response. Past practice in economics has been to try to use multiple regression models to adjust away the differences in observed characteristics, while also testing for sample selection bias. The Propensity Score approach, which is widely applied in epidemiology and related fields, focuses on the idea that "matching" individuals in the groups should be compared. The appropriate matching measure is usually taken to be the prior probability of programme participation. This paper describes the key ideas of the Propensity Score method and illustrates its application by reanalysis of some Irish data on training courses.

I INTRODUCTION

Application of the direct experimental approach in the economy and society is usually considered unpalatable, or even unethical, even when it would clearly provide the ideal comparison. For example, we would like to assess an active labour market policy — say, training to enhance skills — by drawing a

Paper presented at the Fourteenth Annual Conference of the Irish Economic Association.

*We are grateful to an anonymous referee for his suggestions and to the audiences at the conference and at an earlier ESRI seminar for helpful comments.

large and fully random sample from the relevant population and then randomly assigning individuals to a training (or treatment) group and a control group. Then, although an individual's subsequent response (in terms of employment, earnings, productivity, or whatever) will depend on characteristics like age, education and previous work experience, these factors cancel out of the difference in the averages for the two groups.[1] So the difference

$$\bar{y}_1 - \bar{y}_0, \tag{1}$$

where the over bars denote means, and the subscripts 1 and 0 refer to the treatment and control groups respectively, can be validly interpreted as the effect of the programme or policy. But because allocation to a control group can be seen as disadvantageous, randomisation is frequently unpopular, so that randomised experiments are uncommon. There have been some in the US (for example, LaLonde, 1986), but in Europe they are relatively rare with the notable exception of Dolton and O'Neill (1996).

Evaluations have sometimes been based on the performances of the programme participants only, without employing any control group. But then information has to be sought about how individuals would have tried to improve their positions anyway, or else the programme benefits could be considerably overestimated. An Irish example is provided by Breen and Halpin (1988), who evaluated the FÁS *Enterprise* programme by interviewing a sample of participants and, besides ascertaining how well they had got on, also asked what they would have done had the programme not existed. In that study and, no doubt, in many others, there was simply no other way to proceed. But depending on questions of this nature, with the possibilities for "wisdom by hindsight", is less attractive than comparing with a "control" group, even if the allocation to groups has been non-random.

We have a "natural experiment" when we have observational data on a treatment group and on a control group, but without the deliberate randomisation of individuals to groups.[2] Without randomisation, there may well be substantial differences between groups in the distributions of individual characteristics that

[1] The ideal comparison, though impossible to make, would use the *same* people to compare the effect of participation in the programme with non-participation. In experimental approaches the *causal* effect of a treatment on an individual is defined as the difference in the potential responses when receiving and not receiving the treatment. The average of these differences over the whole population is the parameter of interest. While theoretically a useful concept, it is usually unmeasurable and has to be assumed estimable by the difference between the means of the treatment and control groups, given prior randomisation of individuals to groups.

[2] Some authors limit the term "natural experiments" to cases where there are pre and post measures, but we feel the existence of a control is the defining characteristic of an experiment.

affect response. Sometimes quite simple methods are used to analyse the data. For example, if responses of all individuals are measured at two time points, corresponding to before and after treatment for the treatment group, the "difference of differences" method may be employed. This compares the mean improvement in response for the treatment group with that (if any) for the control group, trusting that taking "before" from "after" cancels out individual characteristics. An example is the Eissa and Liebman (1996) study of labour supply response to earned tax credits.

However, most analyses of "natural experiments" in the social sciences have tried to "correct" for between group differences in the distributions of characteristics (covariates) by estimating a multiple regression model (assuming that y is continuous) of the form

$$y_{Dj} = a + bD + \sum_{k=1}^{k=p} c_k x_{kDj} + u_{Dj}\,. \tag{2}$$

Here the x's are the covariates, u is the disturbance term and D is a dummy variable (equalling 1 for treatment and 0 for control) defining groups of programme participants. Its coefficient, b, is interpreted as the treatment effect adjusted to what it would have been had there been no variation in characteristics between groups. Usually, estimation of Equation (2) is accompanied by a test for sample selection bias on the lines developed by Heckman (1976; 1979). The idea is that there could be an unobserved, or latent, variable w (an individual's deviation from average intelligence, say), which affects y. Now if individuals with positive w (above average ability) opt for the treatment group (training, say), while those with negative w opt for the control group, the treatment effect would be biased upwards. If w was observable, the problem would have to be tackled by adding it as a covariate to Equation (2). Of course, this cannot be done when w is unobserved, but if some rather strong assumptions are made, an appropriate proxy variable can be added instead. The assumptions are that the unobserved w also has a linear regression on the x's with a disturbance term v that, jointly with u, follows a bivariate normal distribution. Then through a probit analysis with D as dependent variable equalling 1 for treatment and the x's as explanatory variables, an "index variable"

$$d_0 + \sum_{k=1}^{k=p} d_k x_{Dkj} = -z_{Dj}$$

can be generated from the probit coefficients (the d's) and the covariates and then a proxy variable λ, often called "Heckman's lambda" is provided by the inverse Mill's ratio

$$\lambda_{Dj} = \frac{\phi\left\{(-1)^{D+1} z_{Dj}\right\}}{\Phi\left\{(-1)^{D} z_{Dj}\right\}}$$

where ϕ is the ordinate (or density) and Φ is the integral (or distribution function) of the Standard Normal. Then selection bias would be indicated by non-zero g in

$$y_{Dj} = a^* + b^*D + \sum_{k=1}^{k=p} c_k^* x_{kDj} + g\lambda_{Dj} + u_{Dj}, \tag{3}$$

and treatment effect estimated[3] by b^* in Equation (3) rather than b in Equation (2).

With this formulation of the selection bias issue, the covariates in Equation (2) are being taken as quite sufficient to solve the problem. They come into Equation (3) twice — explicitly as the x's and implicitly in λ, and so, in theory, selection bias is adjusted for by adding a non-linear function of the x's, λ, to the equation. In practice, of course, λ could be nearly collinear with the covariates, making it difficult to conclude anything at all about selection bias. This near-collinearity is avoidable if some of the coefficients in Equation (2) are known to be zero, so that the corresponding variables can be omitted from the response equation, although they are significant when retained in the probit equation. That is, some covariates can be considered to be "instrumental variables".[4] Some standard textbooks (for example, Johnston and Di Nardo, 1997, p. 450) recommend that Heckman "correction" should only be performed in these circumstances. Indeed, if the response variable is itself binary — for example, if a programme is being evaluated in terms of employment gain — the joint model for response and participation becomes a bivariate probit one and then at least one zero coefficient is essential for any estimation. This situation will arise in our data example in Section IV.

The main purpose of this paper is to present an alternative approach to the analysis of "natural experiments" based on propensity score matching. The Propensity Score method derives from papers by Rosenbaum and Rubin (1983a, 1984) and the approach already dominates in biomedical fields. Many expository papers and reviews have appeared in the biometrical literature, including those

[3] Because some variance heterogeneity has been introduced at the probit analysis stage, OLS, while consistent, is less efficient than the appropriate GLS or Maximum Likelihood solution. However, many econometric computing packages provide these procedures.

[4] Of course, with instrumental variables, we could treat D in Equation (2) as endogenous and estimate by IV, bypassing Equation (3) and avoiding the bivariate normality assumption, but perhaps losing efficiency.

by Drake and Fisher (1995); Rubin (1997); D'Agostino (1998) and Perkins, Tu, Underhill, Zhou and Murray (2000), as well as papers describing various variants on the method and many papers describing applications to specific biomedical observational studies. However, we should first outline why the regression approach can prove inadequate and so why consideration of an alternative approach is warranted.

Even before considering sample selection bias, note that many assumptions are already being built into the regression model (2). For example, the effects of covariates are being assumed linear rather than exhibiting increasing or decreasing returns to scale, interactions between covariates are not allowed for, and covariates are assumed to operate identically in the two groups. Of special relevance is the constancy of treatment effect assumption — it is assumed the treatment effect is the same on all individuals receiving it. It should be said that these restrictions on model (2) have long been recognised in the econometric literature and more general models proposed (for example, Maddala, 1983, p. 261). There are also numerous econometric specification tests, although these are not always powerful in practise, but the point is that some degree of incorrect specification error is not at all unlikely in regression equations. In recent years it has been appreciated that the consequences of misspecifications on the estimate of the treatment effect, b, are far more serious when the distributions of covariate values differ greatly between groups than when they do not. For example, Rubin (1997), has described the misleading results that can then be obtained from regression analysis.

As regards Heckman correction for sample selection bias, the validity of the approach is heavily dependent on the postulated selection bias process corresponding reasonably to reality and on the bivariate normality assumption about y and w. Sample selection biases can easily be visualised as operating in a much more complex way than the scenario of people of high ability opting for training and those of low ability opting for the control. There could be selection biases originating with the programme administrators and perhaps interacting with additional selection bias associated with individuals' abilities. There could be selection effects at various stages of programmes — at recruitment, at separation into treatment and perhaps through selective dropout. Sometimes, as in our example in Section IV, the control group may be the artefact of the researchers, composed of individuals selected years later, but felt to have been reasonably comparable to the treatment group at the time of treatment, and it may not be really plausible to think of such individuals as selecting themselves into a control group. Furthermore, the bivariate normality assumption about w and y is sensitive to specification errors in the response equation of the sort already described. For example, when the response equation should have contained some non-linear components, it is known that the λ_{ij} term in Equation

(3) can pick these up and suggest there is selection bias even if none exists. Consequently, there have been criticisms of Heckman's approach in the econometric and statistical literature by, for example, Goldberger (1983); Little (1985) and Holland (1989).

Of course, Heckman's "lambda correction" method was just (an early) one of his approachs to the problem of sample selection bias. Heckman and Robb (1986) and Heckman and MaCurdy (1986) considered IV estimation (as in footnote 4) and other alternatives. Heckman and Holtz (1989), emphasised that the correct choice of selection bias adjustment procedure depended on the source of the selection bias (see, especially, the reply to Holland (1989)). Heckman (1990), like Newey, Powell and Walker (1990), moved towars non- or semi-parametric approaches — a direction continued in some of his more recent papers, which will be mentioned later. However, it is the "Heckman lambda" procedure that is still in textbooks and econometric software packages and that has been, and is, frequently employed in evaluation. Irish examples include Breen (1986); Breen (1991); Callan and Reilly (1993); Breen, Hannan and O'Leary (1995); O'Connell and Lyons (1995); O'Connell and McGinnity (1997) and Doris (1998).

II THE PROPENSITY SCORE APPROACH

The approach depends on the idea of "matching" individuals from the treatment and control groups. Cochran (1965; 1968) gave the example of mortality rates for US smokers being lower, on average, than for non-smokers — the reason being that smokers were younger, on average, than non-smokers. When groups of smokers and non-smokers of equal ages were compared, the mortality rates were always higher for smokers. Cochran advocated seeking causative effects from observational data by matching individuals from the treatment and control groups using all the covariates. If no important covariate has been omitted, it seems plausible to suppose that the difference between the responses of two such matching individuals, one receiving the treatment and the other the control, is the treatment effect plus a random element. Then averaging over the set of differences estimates the treatment effect. Cochran showed that perfect matching, in terms of exact equality of continuous covariates, is unnecessary and that matching on intervals can work well. Nonetheless, the method will meet trouble if there are a lot of covariates, because the number of matching cells increases exponentially with the number of covariates and cells could quickly become empty of treatment individuals, or control cases, or both. That difficulty could be overcome, however, if all covariates could somehow be combined into a single efficient "balancing score". Several ways of constructing such a balancing score have been proposed, but Rosenbaum and Rubin's (1983a) "Propensity Score" approach is overwhelmingly the most popular.

The propensity score for an individual is the *a priori* probability (which is a function of the covariate values) that the individual is in the treatment group. Rosenbaum and Rubin show that if we consider two sets of people, one set in the treatment group and one in the control group, with the same value of the propensity score, then the two sets have the same distributions of covariates. They gave a rigorous version of the intuitive argument that follows. If two individuals, one in the treatment group and one in the control group, have the same propensity score, their subsequent "allocation" to treatment or control can be regarded as if it was random. The difference in their responses is the treatment effect plus a random element and averaging over the set of such differences estimates the treatment effect. So although individuals have not actually been randomly allocated to treatments, the fact that overall distributions of covariates differ between the groups is "ignorable", given matching on the propensity score.

A Propensity Score analysis commences with estimation, by probit or logit, of a treatment assignment equation, where all known covariates affecting assignment and response are included as explanatory variables and the observed "dependent" variable is D=1 for an individual in the treatment group and D=0 for someone in the control group. Then propensity scores are calculated for all individuals and some matching process is implemented. The most commonly employed is stratification of the propensity score distribution by quintiles or sextiles — the "binning" procedure. Then the distributions of covariates for treatment and control within each subclass are compared and, if they still differ appreciably, the assignment equation is further developed. For example, if a particular covariate still differs between groups within subclasses, the assignment model could be modified by trying powers of the covariate and its interactions with other variables. It is important to search for a good model for participation, but "good" means achieving balance of mean propensity scores and of covariates within bins, rather than emphasising the statistical significance of the coefficients in the participation equation. A few redundant covariates (in the sense of not having statistically significant coefficients) are no harm. Indeed sometimes, as will be returned to, we may not wish to include a statistically significant variable. When a satisfactory model is arrived at, the treatment versus control effect on the response variable within subclass i is just the difference in means $\bar{y}_{1i} - \bar{y}_{0i}$, if the response is continuous, or a difference in proportions $\hat{p}_{1i} - \hat{p}_{0i}$, if the response variable is qualitative. Then the overall measure of treatment effect can be taken as simply

$$\frac{1}{s}\Sigma\left(\bar{y}_{1i} - \bar{y}_{0i}\right), \tag{4}$$

where s is the number of bins, or strata, in which there are both treatment and control units and the summation is over these strata. The standard error is

$$\frac{1}{s}\sqrt{\Sigma\left\{\frac{\hat{\sigma}^2_{1i}}{n_{1i}}+\frac{\hat{\sigma}^2_{2i}}{n_{2i}}\right\}}, \tag{5}$$

where the $\hat{\sigma}^2_{1i}$ and $\hat{\sigma}^2_{0i}$ are the within group and within stratum i variances among the n_{1i} treated individuals and the n_{0i} controls. For example,

$$\hat{\sigma}^2_{1i}=\frac{1}{n_{1i}-1}\sum_j\left(y_{1ij}-\bar{y}_{1i}\right)^2.$$

If the response variable is qualitative, formulae (4) and (5) become

$$\frac{1}{s}\Sigma\left(\hat{p}_{1i}-\hat{p}_{0i}\right) \text{ and } \frac{1}{s}\sqrt{\Sigma\left\{\frac{\hat{p}_{1i}\hat{q}_{1i}}{n_{1i}}+\frac{\hat{p}_{0i}\hat{q}_{0i}}{n_{2i}}\right\}}, \text{ with } q = 1 \pm p.$$

The choice of (4) as an estimate needs some discussion. If the treatment effect is the *same* within all strata (that is, training has the same impact on a low propensity individual as a high one) then (4) is not the *best* estimate, although it is unbiased. The best, in a minimum variance sense, would weight the within strata estimates inversely as their variances (so giving most weight to the most precise estimate). But if we want to permit varying treatment effects across strata, we must weight each stratum contrast in proportion to the stratum's fraction of the population — that is equally, given quintiles or sextiles. The constant treatment effect assumption is also implicit in the regression approach of Equation (2), but is not essential for the Propensity Score approach.

The Propensity Score approach is nonparametric as regards the response variable and is very sparing on assumptions. Nothing has been specified about the actual relationships of the response to the covariates, so avoiding the accumulation of biases due to the combination of model misspecifications and unbalanced covariates, which, as mentioned earlier, can have serious consequences for the regression modelling approach. Even the functional form of the relationship between the response variable and the propensity score is unspecified — we just match on the score. This is also why multicollinearity is not the difficulty it can be in regression analysis and why non-significant covariates in the participation equation are not a problem (assuming we are not very short of data). So the familiar criticisms of data mining, pre-testing etc. in regression models do not really apply. The reduction from multidimensional covariates to a unidimensional propensity score, also makes results much easier to interpret and summarise. This point might seem trivial, but it recurs frequently in the literature (for example, Rubin, 1997; Obenchain and Melfi, 1997; Perkins *et al.*, 2000) comparing the Propensity Scores approach with multiple regression type methods. Indeed, some authors have calculated the propensity score and regressed response on it and on a dummy variable for treatment, just as a device to reduce dimensionality in regression, although this is somewhat contrary to the spirit of the Propensity Score approach in that it involves an assumption of constant treatment effect and a strong assumption about functional form.

The value of the Propensity Score method is that it makes few assumptions and with typical programme evaluation data that is an important factor. Of course, if we know that model (2) is exactly specified, with adequate data for its estimation, standard regression is the best approach, as indeed, given exact compliance with the appropriate assumptions, is the Heckman correction (or at least the maximum likelihood version of it) for selection bias. It is also being assumed here that in programme evaluation the treatment versus control comparison is of paramount interest and that we are not trying to estimate multivariate relationships between the response and covariates. The Propensity Score method is not a general substitute for multivariate econometric methods for estimating relationships. While such relationships are certainly important, the presumption is that they can probably be studied (and possibly have) on more data than arise from a single, perhaps minimally controlled, social experiment.

It is possible that the Propensity Score approach could fail to achieve a comparison of treatment with control. The difference within stratum i, $\bar{y}_{1i} - \bar{y}_{0i}$, obviously presupposes that there are some treated and control individuals present. If there are no representatives of one group, that stratum does not contribute to the comparison. If no stratum contains representatives of both groups, the approach fails, since there is no overlap in the propensity scores. The interpretation is that the characteristics (as measured by covariates) of the two groups are so dissimilar that no meaningful comparison is possible. This is not necessarily a disadvantage of the Propensity Score method relative to the econometric modelling approach. The data deficiencies would feed into and undermine the regression analyses, although the cause of the problem might not seem at all obvious. One of the virtues of the Propensity Score approach is that it reveals just how much of the data truly provide information on the comparison. Dehejia and Wahba (1999), who reanalysed Lalonde's (1986) data, emphasise this point. Rubin (1997), writing in the context of drawing deductions from health care databases, has stressed that the first use of propensity scores should be to decide whether a question of interest can be legitimately addressed to the database at all.

However, there is some possibility that a participation equation could be made “too good” and needlessly reduce the degree of overlap. If we *knew* that a variable actually did not affect the response, we would not care whether it was balanced between treatment and control or not, so there would be no need to match on it from that point of view. Including it in the treatment assignment equation would not matter much if it was of marginal significance, but if it had strong predictive power it could lead to an unnecessary reduction in the overlap of the propensity scores. This would make relative frequencies within some bins more unequal and increase the standard errors, or in an extreme case, lose a stratum through absence of one group. This would not bias the comparison between treatment and control, but it would weaken the power of the test of treatment effect. That suggests “instrumental” variables should be omitted from the participation equation. However, if we believe selection bias effects are present, we might wish to make use of them as will be discussed in the next section. In practice though, it is rarely clear that a variable is truly instrumental and Propensity Score exponents usually seem to include all variables in the participation equation.

As an important, although probably obvious, variation, the y variables in Equation (4) could be replaced by differences of post- and pre-treatment (and control) values if the earlier measurements exist. This would give a matched or Propensity Score “difference of differences analysis”, very analogous to common practice in randomised experiments. For example, it has been traditional in animal growth experiments, to use the difference in mean weight gains (assuming initial weights were recorded prior to treatment application) between the treatment and control groups as the estimate of treatment effect. This is more precise than the difference between mean final weights of treatment and control groups, although , given randomisation, that is also an unbiased estimate. In natural experiments there is the added advantage that the difference of differences may cancel out unobserved covariates and the consequent selection bias effects. The pre measure could alternatively be included in the participation equation to achieve matching (to at least some extent) on these unobserved covariates.

At this point it should be said that there are other matching procedures besides stratifying into quintiles or sextiles. More subclasses could be employed, or bins could be based on ranges, rather than frequencies. Leaving stratification entirely, each treatment individual could be matched with the control individual with the closest propensity score value, or matched to a group of “close” individuals, with decisions on “close” based on “callipers” — pre-selected ranges. But most applications of the Propensity Score methodology use “binning” with a relatively small number of bins because Cochran (1968) showed that stratification into quintiles usually removes 90 per cent of the bias due to differing covariate distributions between treatment and control. Another approach to refinement sometimes employed (see Drake and Fisher, 1995, or Rubin, 1997) is regression following matching on the propensity score. This is because model misspecifications cause minimal biases if the covariate distributions are similar for the treatment and control groups.

Other balancing scores have been suggested in the literature as well as the probit or logit based propensity score, for example, that based on the discriminant function (Rosenbaum and Rubin, 1985) or on the classification tree method. Propensity Score analysis requires reasonably large data sets, but if a data set is very large (which it has not been our good fortune to experience) the non-parametric nature of the matching procedure can be further extended by dispensing with the necessity for a probit or logistic participation equation. If covariates are categorical, there will be grouped observations from which participation probabilities can be directly obtained, while Kernel estimation can achieve the same result for continuous covariates. Of course, if the data set is huge, Cochran's (1965; 1968) ideas of matching directly on all covariates would be feasible.

When there are more than two groups comparisons are made pairwise, with separate derivation of propensity scores for each pair. The need for this can be appreciated by extending (following Rubin, 1997) the smoking example to three groups — non-smokers, cigarette smokers and pipe smokers — and two covariates — age and social class. As before, a mortality measure is the response variable. Suppose non-smokers and cigarette smokers have the same age distributions, but unequal, though overlapping, social class membership. Suppose non-smokers and pipe smokers have the same distributions by social class, but unequal, though overlapping, age distributions. Then for the non-smokers versus cigarette smokers comparison the propensity score matching should "balance" social class differences, while for non- smokers versus pipe smokers it should "balance" age differences. Clearly separate derivations of propensity scores are appropriate.

III DEBATES IN THE RECENT LITERATURE

A quick glance at recent literature can give a first impression of substantial dissent, when there is actually agreement about the desirability of "balancing" covariate distributions through matching and a degree of acceptance that, in many cases, propensity scores should be employed. Heckman, Ichimura, Smith and Todd (1996) said that their data analysis "appeared to provide a strong endorsement for matching on the propensity score". The data analysis in

Heckman, Ichimura and Todd (1997) decomposed bias into three components: non-overlapping support, different distributions of covariates within groups and that due to selection on unobservables. They said that this last component "called selection bias in econometrics" had been emphasised in the previous econometric literature, but was actually smaller in magnitude than the others and they concluded: "Simple balancing of observables … goes a long way towards … effective evaluation….".

This is not to say these authors were uncritical of Propensity Score methodology. Heckman *et al.* (1996) said Propensity Score analysis of survey data could be greatly inferior to a true randomised experiment of equal size, which we could not disagree with. But most researchers do not have that choice and will be selecting between methods of analysing non-randomised data. These papers also raise a point followed up in Heckman, Ichimura and Todd (1998), and Heckman, Ichimura, Smith and Todd (1998) — the desirability of matching need not imply that propensity scores are the only, or indeed the best, way to achieve the matching. Their arguments turn on the point that the Rosenbaum and Rubin (1983) proof that matching on propensity scores balances all covariates assumed the propensity scores known exactly, rather than estimated. They argue that matching on *estimated* propensity scores may not be as efficient (in the sense of attaining *asymptotic* bounds to precision) as matching on all the covariates. Hahn (1998) similiarly argues that efficient estimation of treatment effect is possible without the propensity score through "nonparametric imputation" and "various nonparametric regression techniques". As remarked in the previous section, matching on all covariates, or using fully non-parametric methods, requires extremely large data sets and in those circumstances it may not be implausible to proceed as they suggest. The original motivation for the propensity score rather than Cochran's (1965; 1968) matching on covariates was practicality, not asymptotic efficiency.

We think the implications of having to estimate propensity scores are more realistically assessed in the context of data sets of a size requiring parametric estimation via probit, or logit, models and with emphasis on finite sample, rather than asymptotic, criteria. There are relevant papers in the biometrical and statistical literature. These include Rosenbaum (1987, section 2.3) who explained why estimated propensity scores do not have to prove inferior in practice to known true values and Drake (1993) who used simulation studies to examine effects of various misspecifications of the participation equation, including omitted variables, incorrect functional form, etc. She found that many misspecifications did not matter much, provided all independently important covariates had been included.

Another issue debated in the literature relates to the relevance of other measures of difference in response between groups, besides mean difference.

This comes up, for example, in Heckman, Smith and Clements (1997), and Imbens and Rubin (1997). If treatment effects vary over individuals, leading to different distributions of the response variable in both groups, perhaps evaluation criteria need to take account of more than just the different means. Again, there has been debate on when the conceptual comparison of "true" interest might be the effect of treatment on the treated rather than the mean effect of treatment on the sampled population versus the mean "effect" of the control on the population. Heckman and Smith (1996) have even argued that then the measure of treatment effect should include some selection "bias" effect arising from a perhaps desirable selection on ability to the treatment group and that a fully controlled, or randomised, experiment would not be a fair evaluation, because it would prevent this. On a related theme, Angrist, Imbens and Rubin (1996) have incorporated "compliers" and "refusers" into the allocation model. Debates on issues like these are fundamental in clarifying what programme evaluation studies should really address, but we do not think they affect the narrower issue of a matching/propensity score approach versus regression analysis.

We do not think the matters so far reviewed constitute impediments to pursuing a Propensity Score approach. However, the question of whether or how adjustments for unobserved covariates can be integrated into the methodology is one of considerable disagreement in the literature. Matching on the propensity score balances over observed covariates. What if an important covariate — one that has a substantial effect on the response *and* whose distribution differs considerably between treatment and control groups — has not been observed? If it is correlated with the observed covariates, matching on the propensity score will also, at least partially, balance the unobserved covariate effect. For this reason the Propensity Score approach stresses searching for and examining the maximum number of observable covariates. Any further reduction of residual unobservable effects would need additional assumptions. The "classic" Heckman assumptions of Section I permit the residual effect to be fully picked up by the *observed* covariates through a non-linear function (lambda, itself a function of the propensity score). This is very convenient, but the assumptions are strong. Pre as well as post measures of response, permitting a Propensity Score analysis of differences, or integration of the pre measure into the propensity score, would be another mechanism. Here the crucial assumption that the "pre" measures are genuinely free of treatment effect is usually plausible, provided we have a pre measure to commence with, which we often will not.

True instrumental variables could be used to offset effects of unobservables and IV methods have been employed within a Propensity Score framework (for example, Angrist, 1995). However, some authors are very sceptical about how validly instrumental many candidate variables actually are. The choice and framing of assumptions justifying IV estimation, as presented in some econometric papers on programme evaluation and selection, have been subject to critical scrutiny by such authors as Little (1985); Angrist (1995) and Angrist *et al.* (1996), although these criticisms have been disputed.[5] Overall, recent work on IV stresses the importance of ensuring the validity of instruments, that is, their unrelatedness (given other covariates) to the response variable, although those that survive scrutiny are often found much less related to treatment allocation than might be desired (for example, Imbens and Rubin, 1997).

Disagreement is not about possible effects of unobservables having matched on observables,[6] so much as whether an attempted cure via postulated IV's may not be worse than the problem. Heckman, Ichimura and Todd (1998) argue that economists' knowledge of "zero exclusions" or IV's can justify their integration into analyses. Other authors believe that substantial biases are introduced by inexact zero restriction assumptions. Oberchain and Melfi (1997) found models using IV's "frustratingly sensitive to the validity of … underlying assumptions". Little and Rubin (1999) argue that with IV's introduced biases can easily exceed the residual biases from unbalanced unobservables, especially if there has been a comprehensive search for observable covariates. They also cite supporting evidence from large studies (for example, Rubin, Stern and Vehovar, 1995), which, while not programme evaluations, they would consider as being from conceptually highly related contexts. In this literature, elimination of residual biases features less than suggestions for tests (such as, Rosenbaum, 1984; 1989) and sensitivity analyses (such as Rosenbaum and Rubin, 1983b).

So on this issue of how to deal with selection, or balancing, on unobservables, there is still disagreement in the literature, which may take some time to resolve. But it should be noted that the arguments are being conducted in a framework that has taken matching on observables as appropriate. There has been a move away from the parametric analyses represented by models (2) and (3) and it seems unlikely there will be a return to them.

[5] There have even been disagreements over precisely what assumption is implied by an IV. Angrist *et al.* (1996) state that a valid instrumental variable z has to be *independent* of the response variable y given the other covariates. Heckman (1996) said the weaker assumption of z *uncorrelated* (mean independence) with the disturbance term is sufficient, but Angrist *et al.* replied that even if, without independence, there was no correlation with y as the response variable, there would be if log y (or another function) replaced y, as it frequently does in economics. This argument appears elsewhere in the literature, too, for example, in Imbens and Rubin (1997).

[6] Heckman *et al.* (1997) and Heckman *et al.* (1998) used very large data sets to test for the existence of residual selection bias (due to unobservables) following matching for observables. They found evidence of such effects and although the magnitudes were small relative to the biases eliminated through matching, they were still appreciable.

IV EXPOSITORY ANALYSIS OF IRISH TRAINING PROGRAMME DATA

O'Connell and McGinnity (1997) examined the effectiveness of Irish educational training and employment schemes conducted by the State Training Authority (FÁS) and the Department of Education. In mid-1994 they interviewed a large sample (4,600) of individuals who had exited training courses between April and July 1992 and ascertained their pre- and post-training labour market experiences, as well as information on their education, family backgrounds and social circumstances. The courses fell into several categories as regards the type of training and only one category — general training — will be considered here. This is adequate given the expository context of this paper, although the same approaches could be applied to the other course categories. General training courses provided instruction in a range of basic skills and were mainly intended for people with relatively poor educational qualifications.

O'Connell and McGinnity (1997) constructed a control group by selecting suitable people from The Economic and Social Research Institute's long running School Leavers' Survey. This survey takes annual samples of school leavers and follows the cohorts over subsequent years. The criteria for selection were that individuals had left school between 1990 and 1992, were unemployed and in the labour market at the same time as trainees were exiting programmes, and had *not* participated in training courses themselves. They were also interviewed in mid-1994. Clearly these people were relatively young and, to avoid an obvious source of comparison bias, O'Connell and McGinnity excluded all trainees aged over 23 from the analysis. Nonetheless, there were considerable differences between the treatment and control groups in some other possibly relevant characteristics. Table 1 compares the training and control groups in terms of these characteristics, or covariates, showing mean values for continuous covariates and percentages for categorical characteristics and indicating statistically significant differences between groups. Note, in particular, the considerable differences in educational attainments. On average at least, the control group are more advantaged in terms of education and other socio-economic characteristics.

While age was not detected as statistically significant by a t-test, which is not surprising given the exclusions mentioned earlier, this just shows that mean ages did not differ between groups. School leavers were actually a more homogenous group as regards age, while the frequencies of the relatively younger and the relatively older were greater (statistically significantly so) in the training group. So two dummy variables were created corresponding to under 17 years and over 19 years of age.

Table 1: *Analysis of Covariates by Treatment and Control*

	Treatment *Mean*	*Control* *Mean*	*Tests for Difference* *T-Test*
Age (Years)	18.46	18.44	0.17
Duration of Unemployment (Months)	4.54	3.27	3.20**
	%	*%*	*Chi-Square*
Female	43.77	47.97	1.35
No Qualifications	33.91	4.94	79.20***
Junior Certificate	37.36	27.16	8.55**
Leaving Certificate	25.77	65.84	131.96***
Third Level Education	2.96	2.06	0.57
Respondent had never worked	67.92	90.24	47.94***
Father in Employment	42.82	55.00	11.04**
Mother in Employment	16.11	14.58	0.32
Father Employed at School Stage	54.18	65.13	8.93**
Mother Employed at School Stage	7.76	6.22	0.64
Fathers Social Class:			
Professional	11.64	16.45	3.66*
Non-manual Skilled	44.25	47.19	0.62
Semi-unskilled/Manual	44.11	36.36	4.33*

Significance of P at the following levels: *p<.05, **p<.01, ***p<.001.

To implement the Propensity Score approach, we sought logistic models for the probability of participation using combinations of the variables listed in Table 1 and their interactions. The choice of model was partly on standard statistical criteria — the significance of the coefficients (jointly as well as individually), the fit of the model and the retention of observations (some covariates had many missing values). But the procedure outlined in Section II was also followed and models were assessed on how well they achieved within stratum balance of the propensity score and of the covariates. Because there was a considerable degree of colliniarity between variables (including interactions), many models were relatively similar in their overall performance. Although parsimony of model covariates is *not* particularly desirable in modelling the propensity score, for expository purposes we focus on the reasonably simple model shown in Table 2.

Note that there are positive coefficients for the dummy variables corresponding to No Qualifications, or only Junior Certificate and a negative coefficient for the dummy corresponding to Never Having Been Employed. The higher propensity scores and hence the higher probabilities of being in the treatment group are associated with educational disadvantage and lack of work experience. Box plots of the propensity scores produced by this model, categorised by treatment and control, are shown in Figure 1.

Table 2: *Logistic Model for Participation in Training*

Covariate	*Coefficient*	*SE*	*P value*
Constant	1.148	0.308	***
Female versus Male	0.161	0.177	ns
No Qualifications	3.253	0.369	***
Junior Certificate	1.773	0.222	***
Never Worked	–1.894	0.307	***
Under 17	0.75	0.619	ns
Over 19	1.036	0.477	*
Under17*Junior Certificate	–0.786	0.738	ns
Over 19 *No Qualifications	3.701	9.823	ns
Over 19*Junior Certificate	–0.765	0.524	ns
Over 19*Never Worked	0.469	0.521	ns
Model Summary			
–2 Log Likelihood Initial (Constant only)	1136.10		
–2 Log Likelihood Initial with Model Variables	857.7		
R^2 (Nagelkerke)	0.352		
No. of cases	1050		

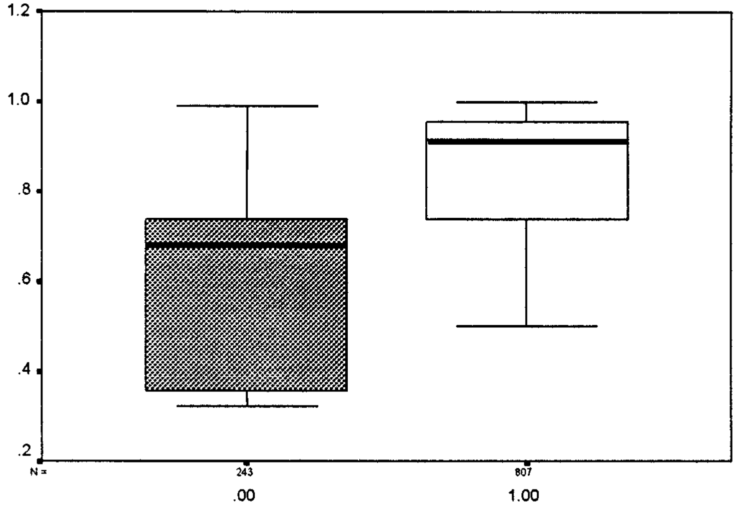

Figure 1: *Within Strata Comparisons of Propensity Scores*

Obviously there is no balance between groups in the distributions of propensity scores, with the training group's values much higher overall. The result of stratifying the distribution into sextiles and comparing distributions within the strata (bins) is shown in Figure 2. Clearly the discrepancies between groups are much reduced within bins. However, the balance has been achieved at the price of different distributions across bins. Table 3 shows that most of the treatment group are in the upper sextiles and most of the control group in the lower sextiles. There are only eight individuals in the control group in bin 5 and six in bin 6.

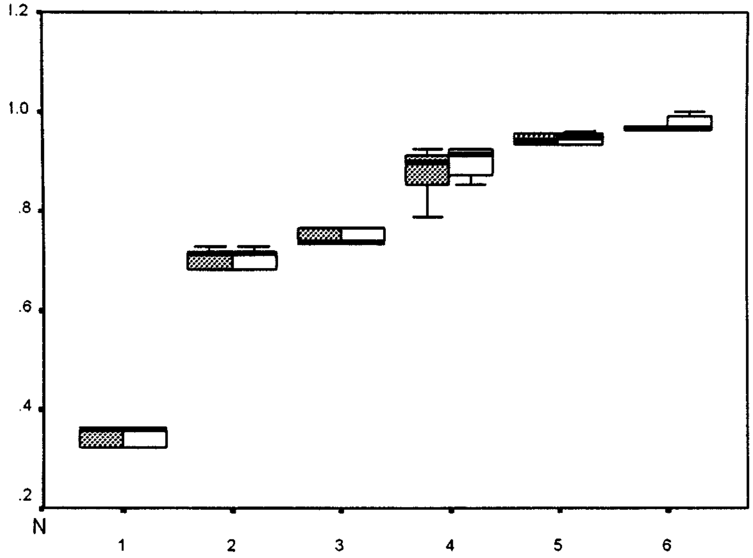

Figure 2: *Within Strata Comparisons of Propensity Scores*

As discussed in previous sections, balancing over propensity scores should balance over the included covariates and this can be checked out. However, this obviously has to be interpreted in the light of the frequency distribution between strata. With the tiny frequencies for the control in bins 5 and 6, the corresponding values are likely to be erratic, but because of correspondingly high standard errors differences from the treatment will not be statistically significant. Table 4 looks at three included covariates — No Qualifications, Junior Certificate and Over 19 years of age and shows how much balance has been achieved. It is probably more interesting to look at the degree of balance achieved over covariates that have *not* been included in the model. So duration of unemployment and (the dummy variable for) the social group semi-and unskilled manual (SUS), both of which are significant in Table 1, have also been included.

Table 3: *Frequencies across Sextiles and Mean Propensity Scores (PS)*

		Bin 1	*Bin 2*	*Bin 3*	*Bin 4*	*Bin 5*	*Bin 6*	*Overall*
Frequencies	Treatment	60	91	157	176	147	176	807
	Control	119	37	54	19	8	6	243
PS Mean	Treatment	0.346	0.704	0.749	0.89	0.947	0.977	0.83
	Control	0.341	0.706	0.747	0.872	0.945	0.97	0.563

Table 4: *Balance Achieved — Means or Proportions and Significance of Differences*

		No Qualifications	*Junior Certificate*	*> 19 Years*	*Duration of Unemployment*	*SUS*
Bin 1	Treatment	0.00	0.00	0.00	2.40	0.42
	Control	0.00	0.00	0.00	1.80	0.28
	Significance	n.s	n.s	n.s	n.s	≈ *
Bin 2	Treatment	0.00	0.24	0.75	6.10	0.34
	Control	0.00	0.22	0.78	4.10	0.45
	Significance	n.s	n.s	n.s	n.s	n.s
Bin 3	Treatment	0.00	0.90	0.00	2.80	0.43
	Control	0.00	0.89	0.00	4.10	0.46
	Significance	n.s	n.s	n.s	≈ *	n.s
Bin 4	Treatment	0.32	0.17	0.55	5.40	0.38
	Control	0.16	0.26	0.63	5.60	0.42
	Significance	n.s	n.s	n.s	n.s	n.s
Bin 5	Treatment	0.35	0.65	0.14	4.20	0.46
	Control	0.50	0.50	0.00	7.90	0.50
	Significance	n.s	n.s	n.s	n.s	n.s
Bin 6	Treatment	0.94	0.06	0.27	5.50	0.56
	Control	0.83	0.17	0.17	6.70	0.50
	Significance	n.s	n.s	n.s	n.s	n.s

Balance has clearly been achieved over the three model covariates. Note how educational levels for both treatment and control groups deteriorate from lower to higher (in terms of propensity score value) sextiles. In Bin 1 everyone had more than just a Junior Certificate. In Bin 6, 94 per cent of the treatment group and 83 per cent of the control group (although this is actually 5 out of 6) had No Qualifications. As regards the covariates that had not been included in the model, it is clear most of their variation is being captured by included variables that are correlated with them. No differences are statistically significant at 5 per cent although for SUS the Bin 1 difference approaches it. Only for Bin 3 did the within stratum difference of unemployment duration approach 5 per cent significance. The difference in Bin 5 looks large, but of course the frequency of the control is low. Including unemployment duration in the propensity score model would have improved it somewhat, but by excluding it and then assessing the balance achieved we have illustrated the previous section's discussion of approaches to, and consequences of, unobserved covariates. Socio-economic variables are often quite correlated with each other and so if substantial efforts have been made to take account of relevant covariates in estimating the propensity score, a considerable amount of balance may also be achieved over an unobserved variable.

Effect of Training on Subsequent Employment

With so few individuals in the control group in bins 5 and 6, these strata cannot separately provide useful information about the treatment effect on the dependent variable especially since the latter is the binary variable — had/had not a job eighteen months after training. However, it seems a pity to have to discard the 176 treatment values, especially since Table 3 and Figure 3 show that overall propensity score values are little different between bins 5 and 6. So in Table 5, which compares the effects of treatment and control on the proportions in employment eighteen months after the completion of the training courses, bins 5 and 6 have been combined. Even so, there are only 14 values for the control, so the within combined stratum comparison will not be at all precise. In the table "Overall" means the comparison *without* any matching on propensity score.

Table 5: *Proportions Employed Eighteen Months Post-training*

		Bin 1	*Bin 2*	*Bin 3*	*Bin 4*	*Bin 5+6*	*Overall*
Frequencies	Treatment	0.50	0.50	0.41	0.37	0.30	0.37
	Control	0.63	0.38	0.39	0.32	0.29	0.49
Significance		n.s	n.s	n.s	n.s	n.s	***

Significance of P at the following levels: *p<.05, **p<.01, ***p<.001.
ns = not significant = not calculated.

Note that the proportions getting employment decline with increasing propensity score. That is perfectly compatible with the fact, remarked earlier, that the control group were more advantaged in terms of educational qualifications etc., so that these characteristics would be negatively related to the probability of being in the training group. But these characteristics would help individuals get employment. So the overall (that is, ignoring the propensity scores) finding that the control group had a highly significant advantage in obtaining employment (49 per cent to 37 per cent), does not demonstrate the failure of training. The important thing is how treatment and control groups compare at similar propensity scores. Within the bins none of the differences are statistically significant at the 5 per cent level, but that for bin 1, is just significant at the 10 per cent level.

The estimation formulae in Section III assumed equal proportions of the data in each stratum and need modification now. For the average treatment effect the formula is now

$$\frac{1}{s_i}\Sigma\left(\hat{p}_{1i} - \hat{p}_{0i}\right),$$

where the within stratum difference in proportions employed is weighted by the stratum's proportion of all individuals. So for sextiles 1 to 4 the s's are 6, and for the combined 5th and 6th, s equals 3. The result is an overall small (.023), but no statistical difference in favour of the training group.

However, there may be a reason to consider bin 1 separately from the others. Table 3 showed that the biggest jump in propensity scores occurs between bins 1 and 2,with the mean for the latter.[7] Perhaps we should accept that general training may not be beneficial to the relatively well educated and advantaged young people who constitute the population of bin 1. Could training be beneficial for the less advantaged (those with higher propensity scores)? Although treatment was not significantly better than control within any one bin, it is worth considering the overall picture across bins 2 to 5+6. A method of assessing if a set of r non-significant tests attain joint significance (Fisher, 1932) is to insert the P_i values from the tests into the formula

$$\chi^2_{2r} = -2\Sigma\log(P_i).$$

For bins 2 to 5+6 this gives a value of 10.2 which, for a chi-squared with 8 degrees of freedom, is still short even of significance. So we cannot assert a definite advantage for training.

Actually, these conclusions are not unlike those drawn by O'Connell and McGinnity (1997) from their "classical" analysis by multiple regression and selection bias testing. Their probit response equation is shown in Table 2.[7] Withholding some covariates from the model was essential, as was mentioned in Section II, if sample selection bias testing is to be conducted via a bivariate probit model.

[7] This is not quite identical to O'Connell and McGinnity (1997) because their analyses were conducted simultaneously for all the categories of training and employment schemes.

Table 6: *Probit Model of Employment after Eighteen Months: General Training Versus Control Group*

	Coefficient	*Standard Error*	*t-ratio*	*P value*
Constant	0.077	0.510	0.150	0.880
General Training	0.024	0.105	0.229	0.819
Female	–0.176	0.085	–2.072	0.038
Age	–0.043	0.031	–1.403	0.161
Junior Certificate	0.565	0.118	4.788	0.000
Leaving Certificate	1.028	0.142	7.253	0.000
Unemployment Duration (pre-programme)	–0.023	0.008	–3.079	0.002
Log Likelihood –631.4 Chi^2 98.3 No. cases 1,011				

There is no statistically significant effect of general training, although some covariates clearly impact on the likelihood of job attainment. Continuing to a test for sample selection bias, the model for the training group could be considered

$$y_j^* = a + \sum_{k=1}^{k=p} c_k x_{kj} + u_j, \qquad (6)$$

where y^* denotes the latent variable underlying the observed binary response variable y and the training effect is considered contained in the intercept, a. The participation equation, which applies to the control group as well as to the training group, is

$$D_J^* = g + \sum_{k=1}^{k=s} h_k x_{kj} + v_j,$$

where D^* is the latent variable underlying the binary variable D, which defines membership of the training or control groups. Sample selection bias exists if a component of v is correlated with u, the disturbance term in Equation (6), for the common training observations. This can be tested for by maximum likelihood estimation of the correlation and seeing if it is significantly different from zero. The procedure can be repeated with the control group replacing the treatment group.[8] The results are in Table 7.

[8] The test could have been applied assuming the same covariate coefficients in the treatment and control groups, but the more general approach was recommended by Maddala (1983, p. 261).

Table 7: *Testing for Sample Selection Bias for the Employment Effects of General Training*

	ρ	*Std. Error*	*t-ratio*	*Significance*
General Training	–0.289	0.264	–1.095	0.273
Control group	–0.281	0.371	–0.756	0.450

The estimates of correlations are not significant and so there seems no reason to modify the conclusion, as drawn from Table 7, that general training does not improve an individual's prospect of gaining employment.

So the two approaches lead to similar findings. But far fewer assumptions were made in the Propensity Score approach and we think it has probed deeper into the data structure. However, the general training versus control comparison was just chosen for expository purposes and no attempt was made to optimise the propensity score estimation. It will be interesting to see if the approach confirms or conflicts with other reported findings from the application of econometric modelling and Heckman correction to the less rudimentary training programmes and to other Irish labour market interventions. We hope, however, that we have adequately conveyed the key ideas of the Propensity Score approach, outlined its scope and illustrated its application.

REFERENCES

ANGRIST, J.D., 1995. "Conditioning on the Probability of Selection to Control Selection Bias", *Technical Working Paper 181*, Washington: National Bureau of Economic Research.

ANGRIST, J.D., and A. KRUEGER, 1991. "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, Vol. 106, pp. 979-1014.

ANGRIST, J.D., G.W. IMBENS, and D.B. RUBIN, 1996. "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, Vol. 91, pp. 444-472. Also — "Rejoinder to Heckman", pp. 468-472.

BREEN, R., 1986. *Subject Availability and Student Performance in the Senior Cycle of Irish Post- Primary Schools*, General Research Series Paper No. 129, Dublin: The Economic and Social Research Institute.

BREEN, R., 1991. "Assessing the Effectiveness of Training and Temporary Employment Schemes: Some Results from the Youth Labour Market", *The Economic and Social Review*, Vol. 22, No. 3, pp. 177-198.

BREEN, R., and B. HALPIN, 1988. *Self-employment and the Unemployed*, General Research Series Paper No. 140, Dublin: The Economic and Social Research Institute.

BREEN, R., D. HANNAN, and R. O'LEARY, 1995. "Returns to Education: Taking Account of Employers Perceptions' and Use of Educational Credentials", *European Socialogical Review*, Vol. 11, pp. 59-73.

CALLAN, T., and B. REILLY, 1993. "Unions and the Wage Distribution in Ireland", *The Economic and Social Review*, Vol. 24, No. 4, pp. 297-312.

COCHRAN, W.G., 1965. "The Planning of Observational Studies of Human Populations" (with discussion), *Journal of the Royal Statistical Society A*, Vol. 128, pp. 234-255.

COCHRAN, W.G., 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies ", *Biometrics*, Vol. 24, pp. 205-213.

D'AGOSTINO, R.B., 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomised Control Group", *Statistics in Medicine*, Vol. 17, 2265-2281.

DEHEJIA, R.H., and S. WAHBA, 1999. "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes", *Journal of the American Statistical Association*, Vol. 94, pp. 1053-1062.

DOLTON, P., and D. O'NEILL, 1996. The Restart Effect at the Return to Full-Time Stable Employment. *Journal of the Royal Statistical Society A*, Vol. 159, pp.275-288.

DORIS, A. ,1998. "Married Women in the Irish Part-time Labour Market", *The Economic and Social Review*, Vol. 29, No. 2, pp. 157-178.

DRAKE, C., 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect", *Biometrics*, Vol. 49, pp. 1231-1236.

DRAKE, C., and L. FISHER, 1995. "Prognostic Models and the Propensity Score", *International Journal of Epidemiology,* Vol. 24, pp.183-187.

EISSA, N., and J.B. LIEBMAN, 1996. "Labour Supply Response to the Earned Income Tax Credit", *Quarterly Journal of Economics*, Vol. 111, pp. 605-637.

FISHER, R.A., 1932. *Statistical Methods for Research Workers*, 4th edn, London: Oliver and Boyd.

GOLDBERGER, A.S., 1983. "Abnormal Selection Bias", in S. Karlin, T. Amemiya and L.A. Goodman (eds.), *Studies in Econometrics, Time Series and Multivariate Statistics*, New York: Academic Press, pp. 67-84.

HAHN, J., 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, Vol. 66, pp. 315-331.

HECKMAN, J.J., 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection Bias and Limited Dependent Variables and a Simple Estimator of such Models", *Annals of Economic and Social Measurement*, Vol. 5, pp. 475-492.

HECKMAN, J.J., 1979. "Sample Selection Bias as a Specification Error", *Econometrica*, Vol. 47, pp. 153-161.

HECKMAN, J.J., 1989. "Rejoinder to Holland", *Journal of the American Statistical Association*, Vol. 84, pp. 878-880.

HECKMAN, J.J., 1990. "Varieties of Selection Bias", *American Economic Review*, Vol. 80, pp. 313- 318.

HECKMAN, J.J., 1996. "Comment on Angrist, J.D., G.W. Imbens and D.B. Rubin", *Journal of the American Statistical Association*, Vol. 91, pp. 459-462.

HECKMAN, J.J., and V.J. HOLTZ, 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: the Case of Manpower Training", *Journal of the American Statistical Association*, Vol. 84, pp. 862-874.

HECKMAN, J.J., and T.E. MACURDY, 1986. "Labour Econometrics", in Z. Griliches, and M.D. Intriligator (eds.), *Handbook of Econometrics* Vol. 3, Amsterdam: North-Holland, pp. 1917-1977.

HECKMAN, J.J., and R. ROBB, 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes", in H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples,* Berlin: Springer-Verlag, pp. 63-107.

HECKMAN, J.J., and J.A. SMITH, 1996. "Experimental and Non-experimental Evaluation", in G. Schmid, J. O Reilly and K. Schomann (eds.), *International Handbook of Labour Market Policy and Evaluation*, Cheltenham: Edward Elgar, pp. 37-88.

HECKMAN, J.J., H. ICHIMURA, and P. TODD, 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies,* Vol. 64, pp. 605-654.

HECKMAN, J.J., H. ICHIMURA, and P. TODD, 1998. "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies,* Vol. 65, pp. 261-294.

HECKMAN, J.J., J. SMITH, and N. CLEMENTS, 1997. "Making the Most of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", *Review of Economic Studies*, Vol. 64, pp. 487-535.

HECKMAN, J.J., H. ICHIMURA, J. SMITH, and P. TODD, 1996. "Sources of Selection Bias in Evaluating Social Programs: an Interpretation of Conventional Measures and Evidence on the Effectivness of Matching as a Program Evaluation Method", *Proceedings of the National Academy of Sciences USA*, Vol. 93, pp. 13416-13420.

HECKMAN, J.J., H. ICHIMURA, J. SMITH, and P. TODD, 1998. "Characterising Selection Bias Using Experimental Data", *Econometrica*, Vol. 66, pp. 1017-1098.

HOLLAND, P.W., 1989. "Comment on Heckman, J.J. and V.J. Holtz", *Journal of the American Statistical Association*, Vol. 84, pp. 875-877.

IMBENS, G.W., and D.B. RUBIN, 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models", *Review of Economic Studies,* Vol. 64, pp. 555-574.

JOHNSTON, J., and J. DiNARDO, 1997. *Econometric Methods,* 4th ed., New York: McGraw Hill.

LALONDE, R., 1986. "Evaluating the Econometric Evaluations of Training Programmes with Experimental Data", *American Economic Review*, Vol. 76, pp. 604-620.

LITTLE, R., 1985. "A Note about Models for Selectivity Bias", *Econometrica*, Vol. 53, pp. 1469-1474.

LITTLE, R., and D.B. RUBIN, 1999. "Comment on Scharfstein, Rotnitzky and Robins", *Journal of the American Statistical Association*, Vol. 94, pp. 1130-1132.

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

NEWEY, W.K., J.L. POWELL, and J. WALKER, 1990. "Semi-parametric Estimation of Selection Models: Some Empirical Results", *American Economic Review*, Vol. 80, pp. 324-328.

OBENCHAIN, R.L., and C.A. MELFI, 1997. "Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies", *American Statistical Association's 1997 Proceedings, Biopharmacology Section*, pp. 297-306.

O'CONNELL, P.J., and M. LYONS, 1995. *Enterprise-Related Training and State Policy in Ireland: The Training Support Scheme,* Policy Research Series Paper No. 25, Dublin: The Economic and Social Research Institute.

O'CONNELL, P.J., and F. McGINNITY, 1997. *Working Schemes? Active Labour Market Policy in Ireland,* Aldershot: Ashgate.

PERKINS, S.M., W. TU, M.G. UNDERHILL, X-H ZHOU, and M.D. MURRAY, 2000. "The Use of Propensity Scores in Pharmacoepidemiologic Research", *Pharmacoepidemiology and Drug Safety*, Vol. 9, pp. 93-101.

ROSENBAUM, P.R., 1984. "From Association to Causation in Observational Studies: the Role of Tests of Strongly Ignorable Treatment Assignment", *Journal of the American Statistical Association,* Vol. 79, pp. 41-47.

ROSENBAUM, P.R., 1987. "Model-Based Direct Adjustment", *Journal of the American Statistical Association,* Vol. 82*,* pp. 387-394.

ROSENBAUM, P.R., 1989. "The Role of Known Effects in Observational Studies", *Biometrics,* Vol. 45, pp. 557-569.

ROSENBAUM, P.R., and D.B. RUBIN, 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika,* Vol. 70, pp. 41-55.

ROSENBAUM, P.R., and D.B. RUBIN, 1983b. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome", *Journal of the Royal Statistical Society B,* Vol. 45, pp. 212-218.

ROSENBAUM, P.R., and D.B. RUBIN, 1984. "Reducing Bias in Observational Studies using Subclassification on the Propensity Score", *Journal of the American Statistical Association,* Vol. 79, pp. 516-524.

ROSENBAUM, P.R., and D.B. RUBIN, 1985. "Constructing a Control Group Using a Multivariate Matched Sampling Method that Incorporates the Propensity Score", *The American Statistician*, Vol. 39, pp. 33-38.

RUBIN, D.B., 1997. "Estimating Causal Effects from Large Data Sets using Propensity Scores", *Annals of Internal Medicine*, Vol. 127, pp. 757-763.

RUBIN, D.B., H. STERN, and V. VEHOVAR, 1995. "Handing 'Don't Know' Survey Responses; the Case of the Slovenian Plebiscite", *Journal of the American Statistical Association,* Vol. 90, pp. 822- 828.